DOCUMENT RESUME

ED 266 172 TM 860 116

TITLE Resource Papers and Technical Reports. Research into Practice Project.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Nov 85
GRANT NIE-G-83-0001
NOTE 174p.; For individual reports see TM 860 117-119.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC07 Plus Postage.
DESCRIPTORS Accountability; Criterion Referenced Tests; *Educational Assessment; *Educational Technology; Elementary Secondary Education; Error of Measurement; *Evaluation Methods; Evaluation Needs; Higher Education; Models; Norm Referenced Tests; *Outcomes of Education; Program Evaluation; *Questionnaires; Self Evaluation (Individuals); Student Characteristics; Surveys; *Validity

ABSTRACT

This document contains three papers developed by the Center for the Study of Evaluation's (CSE's) Research into Practice Project. The first paper, "A Process for Designing and Implementing a Dual Purpose Evaluation System," by Pamela Aschbacher and James Burry, provides a model for evaluating programs for two purposes simultaneously: (1) program improvement; and (2) policymaking. While this paper was written to answer needs of individuals interested in educational evaluation, it can also provide formative information for local program managers and serve the accountability and reporting needs of a state legislature, district office, or other policymaking body. The second paper, "The Credibility of Student Self-Reports," by C. Robert Pace et al., demonstrates that there are many ways to confirm the accuracy, reliability, and validity of student self reports. Part 1 summarizes highlights from the literature and adds comments from the author's research; Part 2 reports on three questionnaires. The third paper "Assessing Instructionl Outcomes," by Eva L. Baker and Harold F. O'Neil, Jr., presents a discussion of outcome assessment that puts into context the evolution of measurement from its beginnings up to its present state. It looks at commonly used psychometric measures, including criterion-referenced, norm-referenced and domain-referenced tests. A special model of evaluation, designed to be adapted especially to the problem of new technologies, is offered. (LMO)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Center for the Study of Evaluation

UCLA Graduate School of Education
Los Angeles, California 90024

DELIVERABLE - NOVEMBER 1985
RESEARCH INTO PRACTICE PROJECT

"Resource Papers and Technical Reports"

Project Director
Joan L. Herman

ED266172

TM 860 116

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

C. Griffith

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

DELIVERABLE - NOVEMBER 1985
RESEARCH INTO PRACTICE PROJECT

"Resource Papers and Technical Reports"

Project Director
Joan L. Herman

Grant Number
NIE-G-83-0001

Center for the Study of Evaluation
Graduate School of Education
University of California, Los Angeles

This document includes resource papers and techical reports developed for the Research Into Practice Project.

Included are:

"A process for designing and implementing a dual purpose evaluation system"
"The credibility of student self-reports"
"Assessing instructional outcomes"

The project presented or reported herein was partially supported by grant number NIE-G-83-0001 from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

A PROCESS FOR DESIGNING AND IMPLEMENTING
A DUAL PURPOSE EVALUATION SYSTEM

CSE RESOURCE PAPER NO. 7

1985

Pamela Aschbacher, Center for the Study of Evaluation
University of California, Los Angeles

James Burry, Center for the Study of Evaluation
University of California, Los Angeles

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California, Los Angeles

TABLE OF CONTENTS

PAGE

INTRODUCTION 1

Purpose and Approach 1

Potential Users 3

Structure of the Resource Paper 4

DUAL PURPOSE EVALUATION: A MODEL PROCESS

Step 1. Identify Evaluation Users and Their Needs 5

Step 2. Form Task Force 6

Step 3. Reach Consensus on Information Needs 10

Step 4. Specify Information Base and Develop Measures 12

Step 5. Develop Procedures for Data Collection & Report of Results 16

Step 6. Pilot and Refine Measures and Procedures 21

Step 7. Collect Information 23

Step 8. Prepare Reports 23

SUMMARY OF CRITICAL CONCERNS 24

CONCLUSION 26

BIBLIOGRAPHY 28

APPENDIX A: Example of Standard Form for Description of Program and Participants

APPENDIX B: Example of Open-Ended Questionnaire for Qualitative Data

APPENDIX C: Sample Checklist for Site Visit Team Leader

APPENDIX D: Sample Site Visit Schedule

INTRODUCTION

Purpose and Approach

This resource paper offers a guide for a dual purpose evaluation plan that can provide formative information for local program managers and simultaneously serve the accountability and reporting needs of a state legislature, district office, or other policymaking body. While the primary audience for this paper is expected to be those interested in educational evaluation, and the examples provided will be from this field, we hope that this paper will also prove interesting to those involved in evaluation of non-educational programs. The approach presented here reflects findings from several years of CSE research on a formative orientation to evaluation, on qualitative methodology, and on strategies for maximizing the utilization of evaluation efforts.

There are three principles that guide our approach. First and probably most important, we believe that evaluation should serve the needs of a multiplicity of users. Teachers, counselors, and program directors continually want evaluation information to refine and improve the program with which they work. In addition, policymakers for such programs (be they at the district, state or federal level) also need information to guide their policy- and decisionmaking. It is clear that these two demands---the top-down demand for broad-level accountability (to improve management and to elevate standards of excellence) and the bottom-up demand for adaptive, sensitive information to be useful at the local level--- push in different and not totally compatible directions. Because it has usually been assumed that these different decisions require different types of data, separate evaluations have usually been conducted to meet these two needs. If we consider the result when these evaluations are added to all the other data collection efforts that occur (e.g., for auditing, accreditation, diagnosis and prescription), we can begin to comprehend why many students, teachers and program administrators feel overwhelmed by evaluation. Furthermore, much of the data collected is of limited utility for them because it often fails to reflect the actual school context and curriculum. They view evaluation as burdensome and intrusive and come to resent the top-down demand for accountability, seeing it as more of a liability than an asset.

Other problems arise when policymakers and program staff have no common basis for their separate inferences about policies and practices. irst, the general intent of educational policy formation is to improve the quality of educational services and to help students attain the highest levels of compentency in school subjects. To accomplish this goal, the policies must be translated into practices that are compatible with the views, needs and capabilities of teachers and students. Second, local programs alone do not necessarily have sufficient resources to solve their problems. The solutions may require initiative, directions, resources and action at those higher levels charged with responsibility for governance, resource allocation and policy formation. In both cases there is high potential for slippage when the information used to assess quality and formulate policy functions independently from that used to actually teach students.

In summary, the current system of independent evaluations appears both uneconomical of time and effort and ineffective in accomplishing the goals of either end of the educational continuum. A partial solution to this problem, presented here, is a single evaluation effort that uses a common information base to serve both policy and program needs simultaneously.

The second principle guiding our development of this paper is that evaluation, testing and standard setting are endeavors which are partly technical, partly political, and partly social. Technical expertise is essential in measurement development and analysis, to ensure valid and reliable use of results. Social understanding is essential to ensure fairness and utility. Similarly, evaluation questions arise out of people's information requirements, while the design and interpretation of evaluations depend on technical competence. The definition of standards depends on values and consensus; the measurement of their attainment involves technical considerations. Thus it seems crucial that a useful evaluation plan will draw on all these areas of expertise and involve professional evaluators as well as the range of users at both levels of the system.

The final principle is that effective evaluation plans are characterized by several key features:

(a) evaluators who are personally committed to the use of results and who have political sensitivity, credibility, and rapport with users;

(b) users who are also committed to implementing the results of the valuation, open to change, and involved in the process so that it reflects their leadership, expectations, and perceptions of needs and risks;

(c) a setting in which representatives of the major stakeholders agree on the focus of the evaluation and the kinds of information it should produce, and in which local program personnel have the autonomy to act upon the evaluation's findings; and

(d) the evaluation procedure is decision-based, allows purposeful sharing of ideas, and provides timely, specific and relevant information (Burry, Alkin & Ruskus, 1985).

Potential Users

The evaluation process described in this paper reflects the need for a systematized, short term, qualitative evaluation plan that addresses both the desire of local programs for formative, program planning and improvement-oriented information as well as the needs of policymakers for standardized information across multiple sites for the purpose of summative decisionmaking, often officially mandated. This evaluation process is intended to be appropriate in a setting characterized by multiple program sites with similar missions (although a single program site could also make use of this plan) overseen by a separate policymaking body. Such settings might include a foundation or community agency with several programs to oversee or the typical educational settings at every age level and level of organization. Specific examples include: (a) a state office overseeing special admissions programs for minority students at each campus of a statewide community college system, (b) a county agency overseeing child care information and referal systems in several cities, and (c) a school district office overseeing bilingual programs in the elementary schools in the district.

Structure of the Resource Paper

The dual purpose evaluation model presented here is described in eight steps:

1. Identify evaluation users and their needs.
2. Form task force.
3. Reach consensus on information needs.
4. Specify common information base and develop measures.
5. Develop procedures.
6. Pilot and refine measures and procedures.
7. Collect data.
8. Prepare reports for policymakers and program personnel.

Where appropriate, specific examples are provided from our experiences with several evaluation projects. The report concludes with a summary of critical concerns.

DUAL PURPOSE EVALUATION PLAN: A MODEL PROCESS

Step 1. Identify Evaluation Users & Their Needs

This evaluation process is designed to meet the needs of two levels of users: policymakers and program personnel. Each of these has somewhat different needs. The purpose of this first step is to identify the specific policymakers (e.g. members of the school board and state legislators) who are responsible for the programs and to identify those individuals involved on-site with the actual implementation of a program (e.g. faculty, staff, and local administrators).

The policymakers usually need summative indicators reflecting program goals and outcomes across all sites at which a particular type of program is implemented. These indicators will focus on broad patterns and trends emerging across program sites rather than on the program as implemented at an individual site. To obtain this information policymakers need a comparable conceptual scheme and measurement base for each site. Where many sites are involved, they may employ an evaluation cycle in which some percent of the total number of sites is evaluated in a given year and each site participates only once every several years. Thus in a given year, the data is collected from only a sample of the whole system and is aggregated so as to provide a view of the system rather than of the individual program sites monitored that year.

The program personnel, on the other hand, need specific information to pinpoint the strengths and weaknesses of their individual implementation of the program in order to make improvements. They want to assess the extent to which the intended program has actually been implemented and the outcomes that have been achieved, both intended and unintended.

Step 2. Form Task Force

Including representatives of all major users of the evaluation (at both the program and policymaking levels) as members of a task force to oversee the evaluation facilitates both the process and the utilization of results. Many an evaluation has been undermined by evaluation users who failed to understand the intent, felt threatened by the potential for change, had political agendas at variance with the goals of the evaluation, were used to pro forma evaluations in which results were never utilized, or were faced with information that could not be clearly interpretted for action. Involving users early in the process helps avoid these problems. It provides valuable input at the point when it is most useful rather than later, thereby causing duplication of effort. It also gives those participants a sense of ownership of the evaluation which makes it much more likely that the results will actually be used for the intended purposes.

The responsibilities of a task force are:

(a) to define the dual focus of the evaluation and reach consensus on what information is needed in the evaluation,

(b) precisely specify the information base and select measures to collect the information,

(c) plan the data collection and analysis procedures,

(d) pilot test and refine the measures, materials, and procedures,

(e) oversee collection of the information,and

(f) prepare reports for policymakers and for program personnel.

In some cases, the task force may preside over a single evaluation. In other circumstances, the task force may be an ongoing body (whose members may change over time) that oversees all two-tiered evaluations of the target programs. In either case, the task force will consist of the general task force members, the task force director, and one or more evaluation experts.

What kinds of backgrounds and skills are sought in task force members? First, they need to represent the major stakeholders in the evaluation: policymakers and program personnel. Where programs serve a number of special groups or provide several services (e.g. a program for the handicapped may serve those with learning disabilities and those with visual impairments in the same "program"), it is important to include task force members with expertise in each of the subgroups or subprograms.

It is important that the task force not be too large to reach agreement on crucial decisions and not be too small to include the full range of experience, expertise, and questions with which the program evaluation must deal. In the case where a very large number of programs is to be evaluated (e.g., the program for handicapped students exists on each of 106 community college campuses in California), it may be prudent to draw representatives from each of the general geographic areas covered by the programs rather than from each of the individual programs.

It is also useful to consider members' organizational position and level of professional experience. For, in addition to overseeing the evaluation, they are also in large part responsible for communicating (both formally and informally) the worth of the evaluation process and results to others in the field. If they are not supportive participants or effective communicators, success will be much more difficult to attain. The evaluation is much more likely to succeed if task force members are in a position of power within their organization and have sufficient experience and skills to put the information to use. Hence, the program representatives on the task force probably should have some administrative responsibilities as well as some responsibilities for planning or providing services or instruction. Task force members drawn from the policymaking level should probably have fiscal as well as program interests and expertise.

In addition to the above, there is a critical constellation of characteristics that all task force members should have: a positive view of evaluation in general, a flexible view of the program and possible changes which may occur as a result of the evaluation, a perception that the evaluation benefits outweigh the risks, and a commitment to use the results to effect recommended changes (Alkin, 1985). People without evaluation experience may be used to operating from intuition rather than from data, so the duties of the task force may be new and somewhat uncomfortable to them at first. Prior experience in evaluation or research is useful in providing appreciation for the necessary technical aspects of the measures and for the feasibility of the data collection plan. In addition, recommendations on these matters from external evaluators are more likely to be appreciated and incorporated without undue delay or resistance. On the other hand, previous experience with poorly conceived or conducted evaluation efforts or those in which results were not used may have led members to distrust evaluators and to expect that the results of the evaluation will never really be utilized, especially for local program improvement. If several members fall in this category, a good deal of time may be needed to change their attitude.

Who should direct the task force? Considering the dual focus of the evaluation, the director must have several characteristics---social, political, managerial, professional and technical--- that will help ensure the success of the endeavor. It is important that all task force members, especially representatives of program personnel (who commonly feel that evaluation is a top-down imposition), view the director as "one of us." A site director of one of the programs, a peer of the other members, may fill this need well if he or she also possesses other necessary traits. Since the members will undoubtedly have divergent opinions on most issues yet be required to reach consensus, it is imperative that the director have excellent interpersonal and communication skills and have power and political acumen. The director's effectiveness in helping the group reach consensus may reflect his or her ability to use the group dynamics to advantage, for example utilizing peer pressure on members who may create problems or stall group progress. The sheer size of the task also requires that the director have strong organization and management skills.

Due to the nature of the task, the task force director must not only have a strong background in evaluation but also be personally committed to seeing the results put to use. Research and experience suggest that evaluations are more effective when users are included in all aspects of the evaluation (Burry, Alkin, & Ruskus, 1985). This may necessitate that the director tactfully educate some task force members and local program personnel about evaluation or measurement concerns. In order to guide the task force well, the director also should have a good sense about how much information is both necessary and feasible to collect.

During at least the initial evaluation, if not the entire life of the task force, the expertise, guidance and encouragement of an external evaluation expert is likely to be needed. The evaluator must have credibility in the eyes of both the program staff and policymakers. Since the evaluator's role is to provide technical expertise and assist the task force, the evaluator must share many of the director's qualities. As with the director, it is important that the evaluator be personally commited to seeing the results put to use. The evaluation is more likely to be successful if the evaluator enjoys a good rapport with the task force members and adopts a collaborative role, in which he or she views the users as colleagues who can help guide the evaluation and who have legitimate questions that are entitled to the evaluator's attention. The evaluator must be sensitive to the program's political dynamics, background, power or prestige. Group facilitation skills will undoubtedly be useful in resolving conflicts and negotiating issues. The evaluator must have ample experience and understanding of the technical requirements of the sorts of measures to be used as well as a realistic expectation of feasible measurement procedures in the given setting. As the task force and its mission attain comfortable age and experience, the need for an external evaluator may recede.

Step 3. Reach Consensus on Information Needs

The first duty of the task force members is to define the purpose of the evaluation and set the ground rules. Beyond agreeing that the evaluation should provide information to guide both policymaking and program improvement, the task force must deal with what balance between these two needs is to be their goal. Are both to be emphasized equally, or should one be more heavily weighted? How is the information to be used? Will it be tied to an audit or accreditation? Is it necessary or desirable for site-specific information to be kept anonymous at the policymaking level?

The second duty of the task force is to reach consensus on what specific kinds of information need to be collected. All task force members should participate in this process, drawing on input from others at the policymaking office and the individual sites. Users' information needs may be stated in terms of questions to be answered or issues to be addressed. Clarifying why users need the given information will help the group set priorities and be certain that their intentions will really meet the need precisely---not just meet it "almost."

Consensus on the general kind of information that needs to be obtained across all sites will provide a uniform information base for the two evaluation purposes. This uniform information base will be referred to here as the "core" information, indicating that it contains the basic information that is required from each program site. In addition, the task force may decide to allow individual sites to supplement the core information with additional site-specific information to be used in program improvement. If this is the case, the task force needs to reach agreement on the type and amount of information each site can collect to ensure that it will not jeopardize the success of evaluating the core issues.

Consensus on information needs is necessary because it guides all future activities of the task force and data collection efforts, particularly development of the specific measurement instruments used at every program site; hence, real, not just rhetorical, agreement by all task force members is imperative for the evaluation to be successful.

One strategy for reaching consensus is for the task force director to present a fairly comprehensive list of tentative questions or issues to which task force members may add others. Members then rank order the issues by importance. Alternatively, each issue could be rated according to its degree of importance. Results can be tallied and further discussed by the group until wholehearted agreement is reached. A good rule of thumb in reviewing all proposed issues or questions, either for the core or the optional portion of the evaluation, is: How, specifically, will this information be used to improve the program or make policy? Will this piece of information really help us make our decisions? The tendency is to say, "That sounds important; of course we need to know that," without seriously questioning how the information will be summarized and translated into proposals for action.

Continued agreement across time may require periodic discussion and particularly clarification, if new members are added to the task force or new concerns are encountered. Significant changes in the purpose, balance, or issues addressed that occurs at any later point in the process will undoubtedly result in wasted time and effort. Hence, it is important to allow sufficient time during this step for the task force to reach a consensus with which everyone is satisfied before proceeding.

Typically, the core information needs include:

a. Who is involved?---a description of program participants

b. What does the program consist of?---a description of program services and/or instruction (which may compare what the program actually is with what it is intended to be)

c. What are the program's effects?---outcome and experiential data for all participants---students, staff and faculty---including performance and attitude data on both intended and unintended outcomes

d. What are the program's strengths and weaknesses?---conclusions and recommendations based on the other information.

Step 4. Specify Information Base and Develop Measures

After the information needs have been defined by the task force, the next step is to write precise specifications for the core information base and the site-specific issues, creating a sort of map that lays out exactly what information is to be collected and how. Specifications cover the content, type of measure, item format, and respondent or source of information. See Figure 1 for an example.

Figure 1.
Sample Specification

CONTENT	TYPE	FORMAT	RESPONDENT/SOURCE
Overall perception of program's strength: a. ways program has benefitted you b. describe one part that works well	questionnaire	open-ended	25% random sample of all students

These specifications will guide development of the actual measures to be selected or created. Both activities, specifying information and creating or selecting measures, have been included in Step 4 because it is often easier to think of them simultaneously. Many people prefer to think in terms of sample items when considering each of the components to be specified rather than trying to consider them in the abstract. In addition, the process of creating measures often suggests new options for the components that were not previously considered, thus causing modification of the original specifications.

When specifying the information base, it is important to note the variety of measures that may be used, such as questionnaires (open-ended as well as multiple-choice), interviews, observations, inspection of written records, and performance tests. There is also variety in the sources of data, all those who participate in the program, the most important of whom are students, faculty, staff, and administrators. Occasionally there are others affected by the program, such as teachers who work with students after they have passed through the program, from whom the task force may want to collect information. It is usually informative to use triangulation of data---asking for the same information from several different sources. Some types of measures, such as follow-up interviews with students, may best be used by trained data collectors during a visit to the program site. Others may be used by regular program participants during the course of the program activities. The task force will probably want to make full use of all these options.

Although many evaluations in the past have tended to focus on quantitative data and have eschewed using much qualitative data due to logistical problems, we would emphasize the value of both quantitative and qualitative data. However, it is easier to collect and use qualitative data if it can be partly quantified on objective, standard forms. For example, demographics on the participants, a description of program services, and a comparison between the proposed program activities and actual services provided may be gleaned from records, interviews and observations. The information may then be used to fill out standardized, highly structured forms common to all sites. An example of one such form is appended to this report.

Qualitative data on how individuals experience the program may be collected via an open-ended questionnaire intended to allow issues to emerge from comments made. (A sample questionnaire is appended to this report.) This data may then be categorized and tallied on standard forms, thereby quantifying the major concerns expressed. If the source of the information is known, interviewers may follow-up on some of the issues with a sample of the people who made relevant comments.

Student outcome data may often be easily quantified, either through the use of existing tests (e.g. curriculum-based tests) or measures developed by the evaluator and task force specifically for this evaluation.

In many cases the desire to gather more information than is necessary or feasible given constraints of personnel, money and time will remain a problem at this stage too. The task force, therefore, may again have to prioritize the list of proposed test items, interview questions, and so forth, to obtain a reasonable amount to measure the agreed upon information base.

To save time and utilize available expertise, the evaluator can take primary responsibility for developing a tentative plan for the specifications, based on the task force's consensus on needs and priorities set in Step 3, with specific input from both policymakers and program specialists on the task force. The initial plan can be presented to the task force as a whole for review. Then the task force can select and refine those ideas which are most useful and generate additional ideas for measures to answer any remaining questions. This model, in which much of the generation of specifications (and measures) is done by the evaluator rather than by the committee as a whole simplifies the job of the task force and allows the evaluator to build in from the beginning certain necessary characteristics of a good evaluation (e.g. sound technical aspects of the measures and a realistic amount of information to be gathered). At the same time, allowing task force members to refine and add measures gives users a sense of ownership of the measures rather than a sense of imposition by the evaluator.

The final task in Step 4 is to select or create appropriate measures (test items, interview questions, observation protocols, questionnaire items, standardized record forms, and so forth) following the specifications just developed. Again, the major responsibility for development of the measures probably best lies with the evaluation specialist, consulting the task force for specialized input, review and modification. The more precise the specifications drawn up earlier, the more straightforward is the development of measures.

Step 5. Develop Procedures for Data Collection & Report of Results

The next set of responsibilities for the task force is to oversee the development of a standardized system for using the measures developed in Step 4 to collect information and write reports for the individual program sites and for the policymakers. A system that is standardized across sites is most likely to provide the uniform information base desired for this type of evaluation.

The specific subtasks include:

a. specifying who will collect which data and when,
b. training data collectors,
c. orienting program participants to their role in the evaluation,
d. planning the reports.

A. Specifying who collects which data and when?

How the data will be collected is largely a function of what the data will be. If the task force wants to compare program plans with actual implementation and also obtain an in-depth, qualitative picture of participants' views of the program, it will probably be necessary to visit each program site to actually see the program in action and talk directly to participants. "One-shot" data, such as performance tests, can also be collected during this visit if desired. On the other hand, if only one-shot,paper and pencil measures are to be used, a site visit may be unnecessary, and the data may be able to be collected by the local program director at each site. If the task force needs data from two or more points in time, such as pre- and post-instruction measures, a multi-phase data collection process will be necessary.

In most cases, a multi-phase process is most appropriate. First, a portion of the data is collected under the direction of the local program director at the individual site, including pre- and post-instruction measures of student accomplishment, faculty, staff and student impressions of the program via written questionnaires provided by the task force, and demographic characteristics of the students, which is available through school records.

Second, a short (1- to 3-day), well-structured site visit by a small team of trained people examines the actual implementation of the program compared to its planned intentions. The team also personally interviews a sample of the program participants. This approach allows them to follow-up on the information obtained during the first phase of data collection, which the team has examined prior to their visit. During the site visit, team members meet regularly to share impressions and cross-validate their observations.

The use of a site visit team has a couple of advantages. It provides the opportunity to bring together data on the same issues from several different perspectives---a means of data triangulation. In addition, the structured, collaborative approach allows qualitative data to be systematically collected and organized for presentation within a very short period.

Who comprises the team? Effective site visit teams utilized by Alkin & Stecher (1981) and Alkin & Ruskus (1985) consisted of three members: the team leader is a program analyst from the policymaking office; the second member is a program administrator who works at a different site from the one being evaluated; and the third member is either an external consultant or program staff member who has had experience or training in naturalistic observation. This seems to be an appropriate model, although the task force may wish to expand the team or make modifications in the members depending on the size and nature of the programs to be visited.

B. Training data collectors.

Regardless of who collects the data, it is crucial that they be given very specific directions. This is particularly true when a group of people are expected to work together as a team. A series of explicit handbooks drawn up by the task force for the site visit team members and the local program administrator can help assure consistency of data across sites. Such handbooks introduce team members to their purposes and process, describe their specific duties, provide tentative schedules and standard data forms, and provide sample reports to illustrate how data may be analyzed and synthesized into the two reports. Some sample handbook components are appended to this paper.

C. Orienting program participants.

In addition to training provided to the data collectors, there is also a need for orienting the program staff at each of the sites to be evaluated. They need to know quite a bit of information in order to cooperate and fulfill their roles: the purpose of the evaluation, how to administer any measures they may be responsible for, how large a sample to select and how to select it if measures are not being given to everyone, where and when to send the information collected, expect to hear the results. If there is to be a site visit, the program staff need to learn what to expect of the visit and should be given relevant details about scheduling and logistics. They may also be given an opportunity to request that the site visit team investigate certain concerns they may have.

D. Planning the reports.

The task force must also plan how the data will be organized into reports for the two major purposes of this evaluation: policymakers and individual program sites. These reports will probably be somewhat different due to the different needs of each of these users. Individual sites need specific information for program improvement, while policymakers may need only a general picture of the overall program, with less or no detail about individual sites. The consensus on users' information needs, reached in Step 3, should guide the content of the reports. The following outlines are to be suggestive rather than prescriptive. These lists of suggested content should be used by the task force to generate a sample site report, since in our experience, it is difficult to obtain a number of comparable reports when only an outline is provided to the authors. (See for an example Burry, 1985.)

The reports to each of the program sites may follow a common outline, such as the following:

1. An executive summary including the recommendations for specific program site
2. A description of the evaluation purposes and procedures
3. Qualitative report, including:
   the setting, participants, qualitative methods used, positive findings, and areas for improvement
4. Report on program activities, including data on demographics of students, staff allocations, and planned versus implemented objectives (description, acceptability rating, and comments)
5. Appendices with all measures used

The report to the policymakers may include the following types of information:

1. Executive overview of contents
2. Description of evaluation purposes and procedures; demographic characteristics of participating program sites
3. Decription of what the program does: clientele, services, personnel
4. Program strengths (overall); optionally, the most outstanding component of each program evaluated.
5. Areas for program improvement: major themes of findings across sites; optionally, highlight specific program recommendations.
6. Appendices with all measures used

Who is responsible for writing each of these reports? The site reports may be written by the site visit team at the end of their visit. The team members may jointly contribute their insights to a unified description of findings and recommendations for improvement. Responsibility for preparation of the report (including answers to both core and optional questions) probably best rests with the team leader. The team may then personally share the results with local program staff and school administrators to use for program improvement. The qualitative portion of the report may, however, be most easily written in the week following the visit since time will be short during the visit itself. If no site visits are conducted, the task force will have to select members to be responsible for receiving, synthesizing and compiling each site's report. In either case, the task force may want to provide each site report author with the description of the evaluation purposes and general procedures, since these should be common to all sites.

The policymaker report is compiled by the director of the task force by aggregating the data from each of the individual site reports. Details on individual sites may be included or omitted depending on decisions made earlier by the task force. Obviously, this report is done after data from all sites has been collected.

One final note about the set of tasks in Step 5 is critical. Many an evaluator has felt that Murphy's Law must have been created to describe evaluations. The next step, piloting and refining measures and procedures, will be made much easier if ample time is allowed at this stage to anticipate problems and find solutions.

Step 6. Pilot and Refine Measures and Procedures

Prior to actual implementation of the evaluation plan it is important to try out the measures and procedures. Good intentions cannot guarantee effectiveness. It is highly recommended, therefore that procedures be piloted at a site not scheduled to be evaluated at this time so that the results can be used solely to improve the measures and procedures. The more similar the pilot test is to the actual circumstances of intended use, the more likely these efforts are to reveal problems that may arise. Depending on the scope of the evaluation and the extent to which the measures and procedures are untried, it may be prudent to do a two-phase pilot. During the first phase, the emphasis may be on the measures and data collection procedures. In the second phase, revised materials may be tried out and reports written to determine if the entire process were capable of generating user-oriented reports.

There are several problems which may arise and which the task force should be prepared for when piloting measures and procedures. Some participants in the evaluation do not do what they are supposed to, either because they fail to take the evaluation seriously or because they did not understand ahead of time how much time or effort it would take. Data collectors may misunderstand what they are required to do and may contaminate the results. Materials may arrive too late to be of use (e.g. pretests arriving midway through the instructional sequence). The measures may be targeted to the wrong people; they may include confusing and redundant items; and most commonly, the measures will be simply too long to effectively administer given the usual constraints.

In addition to piloting the evaluation procedures, it is also useful to try out the analyses and reporting strategies. These components are just as critical as the others to the overall success of the project. It is entirely possible that a measure has been created that looks good, works well with the people taking it, and yet provides data that is unclear or difficult to interpret. Such a problem may not be recognized until someone is forced to synthesize the findings in a report. In addition, this step will indicate whether adequate reports are likely to be produced with the planned materials, information, and procedures.

Step 7. Collect Information

Once the procedures and materials have been piloted and refined, the real evaluation can begin. At this point the task force's duty is to oversee the implementation of the plans and help to solve additional problems as they may arise.

Evaluations in general and site visits in particular sometimes evoke fear of "final certification" or loss of funding. Results will be more useful if the task force is aware of this possibility and can reassure sites about the real purposes of the evaluation. They can emphasize tha the visit is a chance to supplement other data with interviews and direct observation of the program and that it also provides an opportunity for the program staff to request a closer look at certain elements of the program. Finally, the visit allows for dialogue about findings rather than one-way communication.

Step 8. Prepare Reports

The final duty of the task force is to oversee preparation of site reports and to prepare the final report to the policymakers. If the task force has done a good job of planning the contents of and procedures for writing these reports in Step 5 and has piloted and modified the report writing plans in Step 6, this final step should proceed without difficulty.

SUMMARY OF CRITICAL CONCERNS

1. The task force must agree on a uniform information base to meet the designated needs. In one evaluation new members were added to the task force midway through the process, and these new members questioned decisions that had been made without them. This caused the group to stall a number of times to educate these members or to rehash old ground. In addition, "agreement" seemed to have been reached, only to be rescinded later by some members who changed their minds or who had never wholeheartedly agreed in the first place but had not aired their concerns earlier. Utilize peer pressure by members who are "on track" to persuade others to contribute constructively.

2. The task force must involve the key users of the evaluation results, who should understand that the results are truly intended to be used, not just filed away somewhere. This requires a mind set among local program staff to use data rather than mere intuition for program improvement. Ultimately, the usefulness of data for improving local programs rests with the local users---their inclinations, attitudes and schemas. Are they truly trying to improve the program, or is personal survival their first priority?

3. The measures must be relevant to the agreed upon focus, technically adequate, able to produce meaningful and useful results, and be acceptable

to the users. This will require a delicate balance of input from users and professional evaluators. In a previous evaluation where measures were to be developed primarily by the task force members with guidance from the evaluation expert, the resulting measures were wildly disparate in size, scope and technical properties. Yet the evaluator felt constrained by needing task force members' ownership of the measures. Perhaps such ownership can be obtained through review rather than painstaking development of specifications.

4. Standardization of procedures and careful attention to logistical details will help assure comparable, meaningful information with fewer negative side effects for evaluation participants.

5. Realistic expectations regarding what can be accomplished given the available time, resources and personnel are critical to success. Task force members may be reminded of the importance of a few good measures, done well, with clear implications for action over a large, messy hodgepodge that is ultimately forgotten.

Conclusions

In this paper we have provided a model for evaluating programs for two purposes simultaneously: for program improvement and for policy making. This dual focus approach has two major advantages by providing a common information base for decisions at both ends of the continuum, it helps ensure that these decisions are in harmony with each other. In addition, this approach conserves time and effort of staff and students alike by making one data collection effort serve two purposes.

This user-oriented model reflects many years of evaluation experience and has been used successfully in several educational settings. It relies on a task force of evaluation users, who are open to the possibility of change and are committed to using the results. As a group they must reach consensus n their purposes, desired information base, and procedures, and then see that the plan is carried out.

The model evaluation plan presented here consists of eight steps, as follows:

1. Identify Evaluation Users and Their Needs
2. Form Task Force
3. Reach Consensus on Information Needs
4. Specify Information Base and Develop Measures
5. Develop Procedures for Data Collection and Report of Results
6. Pilot and Refine Measures and Procedures
7. Collect Information
8. Prepare Reports

While this plan was based on evaluation experiences in education, it should also prove useful in other fields in which there is a need to make data-based decisions to improve local programs as well as to create policy that serves a number of such programs.

BIBLIOGRAPHY

Alkin, M.C. (1985). A guide for evaluation decision makers. Beverly Hills, CA: Sage Publications.

Alkin, M.C. & Ruskus, J. (1986) Systematizing short term qualitative evaluation. Paper to be presented at AERA, Spring, 1986.

Burry, J. (1985). Evaluation of martello community college's handicapped students programs and services: a prototype evaluation report. Los Angeles: Center for the Study of Evaluation, UCLA.

Burry, J., Alkin, M.C., & Ruskus, J. (1985). Organizing evaluations for use as a management tool. Studies in Educational Evaluation, 11, 131-157.

Morris, L.L. & Fitz-Gibbon, C.T. (1978) Evaluator's handbook, Beverly Hills, CA: Sage Publications.

-----(1978) How to deal with goals and objectives. Beverly Hills, CA: Sage Publications.

-----(1978) How to design a program evaluation. Beverly Hills, CA: Sage Publications.

-----(1978) How to measure program implementation. Beverly Hills, CA: Sage Publications.

-----(1978) How to measure achievement. Beverly Hills, CA: Sage Publications.

-----(1978) How to measure attitudes. Beverly Hills, CA: Sage Publications.

-----(1978) How to calculate statistics. Beverly Hills, CA: Sage Publications.

-----(1978) How to present an evaluation report. Beverly Hills, CA: Sage Publications.

Appendix A

Sample Standard Form for Quantitative/Qualitative Data

Part IV: Staff Allocations: Indicate only people for whom there is a discrepancy between proposed and actual activities

Name	FTE	Discussion of Discrepancy (Major variances by component #)	Rec #

Appendix B

Confidential Student Questionnaire

Directions: Please answer the following questions. To maintain confidentiality, please seal your completed questionnaire in the attached envelope. Sealed questionnaires will be read by review team and will not be shown to local students, staff, or administrators. Please be candid -- your comments can help to improve the project.

Thank you.

1. In what ways do you feel the project has been of greatest benefit to you?

2. Describe one part of the project here that you feel works particularly well.

Student Questionnaire (Cont'd)

Why do you think it's effective?

3. Describe one part of the project here that is in need of improvement.

Why do you think it needs to be improved (What's wrong with it)?

Please specify your major area of study ______________________________

Please indicate services you have received from the project:
Financial aid ____ Counseling _____ Tutoring ____ Recruitment ____
Employment assistance ____ Childcare ____ Transportation ____

Your name ______________________________________ (Optional)

Appendix C

CHECKLIST

Team Leader

Prior to the Site Visit

- ☐ Identify college to be visited and select individuals to serve as Team Members B and C.
- ☐ Schedule visit and make travel arrangements.
- ☐ Prepare Site Visit Packet.
- ☐ Mail Site Visit Packet and Handbook with supplements and proper site visit schedule to Team Members B and C.
- ☐ Mail to Program Director the "Director's Guide" and multiple copies of Confidential Student and Faculty/Staff Questionnaires, and survey of student goals.
- ☐ Check with Director to assure that student goals survey has been conducted and confidential questionnaires distributed and returned. Remind Director to fill out Summary of Project Accomplishments form and Student Population form which are part of the "Director's Guide."
- ☐ Be certain that Director mails completed confidential questionnaires to Team Member C at least ten days prior to site visit.

During the Site Visit

Coordinate all activities including:

- ☐ Conduct brief team meeting before the visit to get oriented to procedures.
- ☐ Informal Introduction.
- ☐ Planning Meeting (Team Members A, B, and C meet with Director).
- ☐ Assign Interview Tasks (Team Members A and B).
- ☐ Campus Orientation (Team Members A, B, and C tour campus).
- ☐ Gather and Record Data (Team Members A and B jointly interview Director; Team Members A, B, and C separately interview other staff members and students, observe project activities, and review documents).

- ☐ Prepare for exit interview (Team Members A, B, and C prepare recommendations and compose brief description of findings).
- ☐ Conduct Exit Interview (Team Members A, B, and C with college president and administrators, Director and senior project staff.

After the Site Visit

- ☐ Prepare final report.
- ☐ Distribute final report to Director, college president, college administrator who supervises Director, to each of the other team members, and to Task Force Director (for synthesis in Report to Policymakers).

Appendix D

Tentative Schedule

(Two-Day Site Visit)

Time	Activities	Team Members
Evening Prior to Day 1	- Arrive at hotel	A+B+C
	- Team meeting to review site visit schedule and responsibilities	A+B+C
Day 1		
8:30-9:00 am	- INFORMAL INTRODUCTION - Meet with project staff - Explain purpose of site visit - Coffee	A+B+C
9:00-10:00 am	- PLANNING MEETING: - Meet with Program Director to review project data and to determine interview sources	A+B+C
	- ASSIGN TASKS	A+B
10:00-12:00 am	- CAMPUS ORIENTATION: Tour of campus and EOPS facilities. Visit with President or his designee to comment on purpose of of visit. Scheduled meeting and short interview with immediate supervisor of Program Director	A+B+C
12:00-1:00 pm	- Lunch	
1:00-2:30 pm	- GATHER AND RECORD DATA: Interview Director Interview other staff members and students	 A+B C
2:30-5:00 pm	- GATHER AND RECORD DATA: Interview other staff members and students	A+B+C (separately)

Time	Activities	Team Members
8:00-9:30 pm	- Team meeting to review progress, discuss preliminary recommendations, and coordinate activities for Day 2.	A+B+C
Day 2		
8:30-11:00 am	- GATHER AND RECORD DATA: Interview staff members and students, observe project activities, and examine documents	A+B+C (separately)
11:00-12:00 am	- GATHER AND RECORD DATA: Interview Director	A+B
12:00-1:00 pm	- Luncheon meeting among team to monitor status of site visit and plan final activities	A+B+C
1:00-2:30 pm	- GATHER AND RECORD DATA: Final interviews, observations etc.	A+B+C (separately)
2:30-3:30 pm	- Team meeting to prepare final recommendations and summary of findings	A+B+C
3:30-4:30 pm	- EXIT INTERVIEW with college president and administrators, EOPS Director, and senior project staff	A+B+C
4:30 pm	- Depart	

THE CREDIBILITY OF STUDENT SELF-REPORTS

Prepared for the Center for the Study of Evaluation
Graduate School of Education, UCLA

by C. Robert Pace

with the assistance of

Doris Barahona
David Kaplan

November 1985

Acknowledgments

I am indebted to Doris Barahona and David Kaplan for important portions of this report. Both are graduate student research assistants in the Graduate School of Education. Following discussions with Doris Barahona about the sort of internal cross-tabs that might bear on the credibility of student reports to the College Student Experiences questionnaire, she identified a whole array of questions and answers that seemed to be relevant, and then obtained the results via many computer printouts. David Kaplan explored the possible relevance of multivariate statistical analyses for judging the predictive and the construct validity of student self-reports in the College Student Experiences questionnaire.

In the process of thinking about and then producing the present document, I have benefited from discussions with these colleagues and I have welcomed and appreciated their interest.

C. Robert Pace
November 1985

INTRODUCTION

Whenever one presents the results of a questionnaire survey, there is always someone who says "But those are only opinions". If the results come from a survey of students, the put-down response is "But those are only students' opinions", as if, coming from students, the results are even less believable. If the comment comes from someone in the "hard" sciences, it is likely to be "But you only have 'soft' data".

It's interesting that this sort of knee-jerk disbelief does not automatically occur in response to other surveys. The Census Bureau conducts many surveys that ask about people's opinions and plans. There are surveys to estimate consumer confidence which are taken seriously by economists and entrepreneurs. Political opinion surveys are carefully studied by candidates for office. Opinion surveys are an important aspect of market research. There is, of course, a certain skepticism about the credibility of some self-reports to the Internal Revenue Service. But on the whole, opinion polls, survey research, and questionnaires are widely accepted methods of inquiry, and certainly a very significant feature of scholarship in the social sciences.

Opinion polls and attitude surveys, like other inquiries, are subject to errors of measurement. For more than fifty years there has accumulated a very large body of research on possible sources of error, and on ways to estimate reliability and validity. The Public Opinion Quarterly regularly publishes scholarly articles on the methodology of polls and surveys. The major polling agencies are especially sensitive about the accuracy and validity of their reports. Some of the best known survey centers are university-based -- as the National Opinion Research Center at the University of Chicago, and the Institute for Social Research at the University of Michigan.

In higher education, and in education generally, questionnaires are quite common. There has also accumulated over a period of years a body of research on the credibility of students answers to questionnaires. The present report on the credibility of student self-reports is a preliminary document that should, and perhaps may, become a more thorough and scholarly document at some future date. Meanwhile, we aim to present a few highlights from the large literature on measuring attitudes and other subjective phenomena, note some of the accuracy checks that have been made with respect to college student questionnaire responses, and then examine briefly the features of two current questionnaires for entering college students and explore more extensively one current questionnaire for undergraduates to illustrate a variety of reliability and validity estimates that can sometimes be produced to demonstrate the credibility of students answers.

PART 1

ISSUES, ANSWERS, AND ADVICE

The Russell Sage Foundation has recently published a definitive two volume document entitled Surveying Subjective Phenomena, (Turner and Martin, Editors) 1984. For anyone who wishes to review the literature of research on this topic, those two volumes are a fairly complete answer. In addition, the Russell Sage Foundation has also published a book by one of the most highly regarded scholars, Otis Dudley Duncan, Notes on Social Measurement: Historical and Critical, 1984, which deals with the whole domain of counting and classifying demographic and other elements, from antiquity to the present.

In 1976 the College Entrance Examination Board published a monograph by Leonard Baird, Using Self-Reports to Predict Student Performance, which reports much of the evidence from college student surveys about the accuracy of their responses to questionnaire items, as well as their utility for prediction.

Part 1 of this report is not a review of the literature in the usual sense. No attempt is made to cite chapter and verse from dozens of studies. Rather, everything (except as may be subsequently noted) that will be mentioned comes from one or more of the four major sources just cited. What follows, then, is my summary of what I regard as a few highlights from the literature, plus some of my own contributions to that literature over the past 50 years.

Varieties of Self-Reports

Some self-reports merely ask for obvious, easily verifiable information, such as age, sex, marital status. It is a subjective or individual answer to an objective question. At the other end of the spectrum are questions and answers both of which are entirely interpreted by the individual. A good example is the following question: "Taken all together, how would you say things are these days -- would you say that you are very happy, pretty happy, or not too happy?" An example from a survey of college alumni is the following: "What is your present feeling about your college? -- strong attachment to it, pleasantly nostalgic but no strong feeling, more or less neutral, generally negative, thoroughly negative". The meaning of the question and of the response is determined by the respondent, and can be directly known only by the respondent.

In one part of the appendix to Volume 1 of the Russell Sage report there is a "Scheme for classifying survey questions according to their subjective properties" (pages 407-431). The main categories of this scheme illustrate the varieties of self-reports one encounters in surveying subjective phenomena. There are three dimensions. The first is the referent of the question: objective versus subjective events. Objective questions refer to events that can be externally observed. Subjective questions refer to internal conditions, intuitions, beliefs, etc., which are directly knowable only by the individual. The second dimension is the nature of the judgment. Such judgments might involve beliefs, attributions, or valuations, and they involve different intellectual tasks. Simple judgments about the occurrence of events primarily involve recall. Attributions require generalizations and inference. One finds very generalized referents such as "most people", "all in all", "people running the country today", "most faculty members", etc. The interpretation of answers is complex and surely suggests the importance of skepticism. Valuations include questions about preferences, likes and dislikes, approval ratings, attitudes toward people, groups, organizations, policies, subjective sentiments such as confidence ratings, satisfactions, problems and worries. The third dimension is the object of the report: self versus other. Is the respondent being asked to report about himself? If so, do people tend to present themselves in a good light? How do these self-perceptions influence one's perception of others?

These three broad categories, albeit overlapping in some respects, are useful to keep in mind as one examines the content of questionnaires: the referent of the question, the nature of the judgment, and the object of the report.

Errors of Measurement

In questionnaire surveys of college students the chief source of unrepresentative results are the nature and size of the sample, and the proportion of people who return the questionnaire. Students in a large introductory psychology course are often asked or required to respond to some questionnaire. They, of course, are not a representative sample of anything. For relatively small colleges, the best advice is to give the questionnaire to everyone, thus bypassing the sampling problem. In big universities, the task of having all entering freshmen respond to a questionnaire is never successfully completed. If one can get two-thirds or three-fourths of the population one is doing rather well. There are good studies that have obtained data from a broad assortment of students and institutions; but nothing comparable to a national public opinion poll in its representativeness. The more significant problem, however, is in the response rate. Whether questionnaires are distributed via the U.S. Postal Sevice, or whether they are put in a campus mailbox, many are never returned.

In a national questionnaire survey of students and alumni which I carried out in 1969, involving random samples at about 75 colleges and universities, the median response rate to the freshman questionnaire was 80%, for the upperclassmen questionnaires the median response rate was 66%, and for the alumni samples the median response rate was 58%. The questionnaires, each about 16 to 20 pages in length, were attractively designed and printed; most colleges used one followup reminder; and for the alumni samples there were two followup reminders.

Even if one had returns from everyone the basic conclusions would not change significantly; but probably in all questionnaire surveys there is some selectivity or bias among those who respond. In the 1969 study the poorest rates of return from freshmen and upperclassmen came from the large institutions; but in the alumni questionnaire the differences in return rates were not related to size, they were related to institutional selectivity and prestige. In the elite categories, only 2 in 20 (10%) had an alumni response rate of less than 50%; in the middle category scholastically, there were 10 of 39 (26%) with a response rate of less than 50%; and in the least selective category, there were 5 of 15 (33%) with fewer than 50% returns from their alumni.

In two recent questionnaire surveys of UCLA undergraduates, the response rates have been between 45% and 50% There are, of course, ways to increase the rate of return of mailed questionnaires. Unfortunately, for academic researchers, they are very costly and the money is not forthcoming.

Unlike the usual procedure in academic surveys, the national opinion polling agencies collect their data by interviews. The carefully designed stratified area sampling techniques do, in fact, produce reasonably reliable and valid results. The magnitude of non-response is minimal because the interviewers's job is to get everyone who fits the sample specifications.

On several past occasions I have suggested that periodic polls of college students might be very worthwhile. But they would require developing an adequate base for sampling, and this does not now exist. The carefully designed sampling procedures, and the resulting national samples for public opinion polls, are not applicable to the college population.

There are several other aspects to the present topic of measurement error. These relate to the estimation of reliability. Does one get similar answers to the same questions from comparable samples? In a test-retest situation, do people give the same answer the second time that they gave the first time? Do slightly different questions about the same topic result in generally similar responses? Most surveys in social science and in higher education do not report answers to any of these questions, and presumably do not collect evidence about any of these matters. But they should. And at least periodically they have.

In 1948 a 16-page questionnaire was mailed to a sample of Syracuse University alumni. The questionnaire included two types of items which were subsequently readministered to a small sample. The questionnaire contained eleven Activity Scales of eleven items each, labeled Politics, Civic Affairs, Religion, Art, Music, Literature, and Science. The subjects checked each activity they had engaged in during the past year. The scales were Guttman-type scales in that participation in the more dificult activities tended to subsume participation in the easier and more common activities. The score on each scale was simply the number of activities checked. Then there were nine Opinion Scales of six items each, labeled Politics, Civic Relations, Government, the World, Philosophy, Art, Music, Literature, and Science. The statements in the opinion scales were written to reflect basic concepts or generalizations about the topics, generalizations reflecting a consensus of experts in the field, so that it was possible to score each scale simply by counting the number of statements on which one's opinion agreed with the opinions of the experts. Each statement was answered on a five point scale, from Strongly Agree to Strongly Disagree. Six months after the initial sample of 2500 had filled out the questionnaire, a second copy was sent to a small group of 120, receiving 68 in return. The test-retest consistency of scores over this six-month interval was computed. For the Activity Scales, the correlations ranged from .70 to .89, with a median of .83. For the nine Opinion scales the median test-retest correlation was .65, with seven falling between .60 and .70, and two much lower ones of .40 and .31. Consistency of responses was also checked item by item. For the Activity items, the averge percent of identical responses was 85, with a range from 83 to 87. For the Opinion items the average percent of identical responses was 75, with a range from 68 to 84. The above test-retest data were reported in an article by Pace, "Opinion and Action: A Study in Validity of Attitude Measurement", Educational and Psychological Measurement, Vol. 10, No. 3, 1950, pages 411-429.

The ACT Evaluation/Survey Sevice, Users Guide, 1981, reported test-retest results on ACT's Student Opinion Survey for a group of students at one university who responded to the questionnaire a second time approximately two weeks after the initial response. The average percent of identical responses on the two administrations was 98% for demographic background items (age, race, sex, etc.), 90% for other background items such as hours worked per week, occupational plans, etc., and 93% for items about the usage of college programs and services. For "Satisfaction" items (responses on a five-point scale from Very Satisfied to Very Dissatisfied) referring to such matters as academic aspects of the college environment, rules and regulatioms, facilities, college services, etc., the percent of identical item responses was typically about 64%, and the percent of responses within one scale point of the identical response typically about 95%.

In the American Council on Education Research Report, Vol. 7, No. 2, 1972 by Boruch and Creager, entitled Measurement Error in Social and Educational Survey Research, two examples of test-retest comparisons are cited. One example administered a questionnaire twice, with six weeks intervening, to a group of 107 college students. Questions about students previous achievements resulted in 90% to 100% agreement. Answers to other facts -- such as father's education and occupation, high school grades, etc., had agreement percentages from 74% to 92%. Attitudinal items, and questions about future plans typically involved agreement in the 60-70% range. The other example was the readministration of the ACE freshman survey questionnaire to 202 students following an interval of two to three weeks. Test-retest correlations for different types of items were as follows: demographic characteristics, mostly .96 to .99; sources of financial support, mostly .85 to .88; self-reported attributes of parents, mostly .60 to .82; items estimating the chances of future events (such as graduating with honors, joining a fraternity or sorority, failing one or more courses, changing career choice, etc.), mostly .58 to .88 with a median of .78; items about life goals such as the importance of being very well-off financially, raising a family, keeping up with political affairs, helping others in difficulty, mostly from .65 to .87 with a median of .73; attitudes toward the importance of various federal actions such as pollution control, school desegregation, veterans benefits, consumer protection, correlations ranging from .41 to .83 with a median of .63; and items about attitudes toward various campus and social issues such as faculty promotions should be based on student evaluations, marijuana should be legalized, with test-retest correlations ranging from .57 to .88 with a median of .66.

Both the ACT and ACE reports show that the greatest variability in responses are found in relation to questions that are ambiguous, or about topics which students may not have given much prior thought or concern , or about attitudes which are themselves subject to various interpretations. In some cases, the test-retest correlations are low enough to raise doubts about the value of the responses, especially when the test-retest interval is only 2 to 6 weeks. For the more specific items, consistency of responses was quite high.

In public opinion surveys there have been some examples of comparing the results to the same questions when asked by different survey organizations. The closest or most carefully controlled conditions are called tandem surveys. In one such tandem survey, NORC and Roper each drew probability samples and proceded to administer the survey in their customary fashion. This was a survey about public use of and attitudes towards television. Differences in the results were small; but there was a clear effect related to how the organization determined the "don't know" responses. On 52 comparisons, NORC had fewer DKs on 42 items, Roper fewer on 4 items, with no differences on the other items. In another study, a survey about public attitudes and knowledge concerning survey practices, the sample was drawn by the Survey Research Center, and the cases randomly assigned to SRC and Census Bureau interviewers. In general, the results were fairly similar. However the interviewee refusal rate was 6% to the Census Bureau interviewers and 13% to the SRC interviewers.

A summary tabulation reported in Volume 1 of the Russell Sage publication, of 126 instances in which the same questions were asked by different survey measurement programs at about the same time shows that in 45 of the instances there were differences beyond the level typically allowed for sampling error. Such differences could have come from many sources -- context, interviewer effects, training and staff differences, etc. Some of the differences were clearly attributable to how DKs were handled. Variations in practices produce differences in the products; but most of these differences are relatively small. When the conditions are most comparable, as in tandem surveys, the results are highly congruent.

Errors of Substance

Whether people report accurately about their conditions or their behavior is, in one sense, an error of measurement and in another sense an error of substance. In surveys of college students there is a good deal of evidence that self-reports about their school grades, and about prior accomplishments are very accurate. Much of this literature has been summarized by Leonard Baird in the monograph he wrote for the College Board in 1976. Are student's self-reports of their grades accurate? Baird himself found that the correlation between college-reported and student-reported grades was genrally about .87. In a study of self-reported and transcript-reported grades, by Nichols and Holland in 1963 among National Merit Scholars, cited by Baird, the correlation was .96. Maxey and Ormsby in 1971 reported correlations between self-reported and school-reported grades in a sample of nearly 6000 students in 134 schools to be on the average in the mid eighties. They found that 98% of the students' reported grades were accurate within one grade. Baird concludes from many studies that "research accumulated over 30 years, using various methods, in samples of grade school students, high school students, college applicants, junior college students, four-year college students, and professional school students, adds up to one conclusion: students' reports of their grades are about as useable as school-reported grades". (page 8). Moreover, self-reported grades predict future grades as well as or better than college entrance tests of academic ability. It seems fair to conclude that, at least for some kinds of questions, errors of substance in the answers are minimal.

The data from the above studies are a good example of what one can expect when the questions are clear and specific, and when the response options are equally clear and specific. Students know the definition of grades and they know their own grades. Consequently, one can have confidence that the subjects can answer the questions. But in many surveys no such clarity is evident.

Evidence from the larger survey research literature also confirms the accuracy of self-reports about various specific conditions or behavior. For example, correlations between employers records about wages, duties, etc., and application blank work histories were generally .90 or greater. Adults reports of whether they owned their home were 96% accurate, had a valid library card 87% accurate. One needs to be reminded here, that "official records" are not always 100% accurate.

Perhaps one of the most serious errors of substance arises from variations in the content, or wording, of the questions, and from the context in which the questions are used. There are some classic examples of this. The following question was asked in a national sample poll: "Do you think the United States should let Communist newspaper reporters from other countries come in here and send back to their papers the news as they see it?" Half the questionnaires asked this question after another question on whether the Soviet Union should allow in American newspaper reporters; and the other half of the questionnaires asked the questions in the reverse order. When the question about communist reporters was asked first, 55% of the people agreed; but when the question about American reporters was asked first, 75% agreed. Or, consider the following two questions: 1) Do you approve or disapprove of a married woman earning money in business or industry if she has a husband capable of supporting her? (65% of a national sample approved); 2) If there is a limited number of jobs, do you approve or disapprove of a married woman earning money in business or industry when her husband is able to support her? (Only 36% approved!) Here is another example of different answers from slight differences in wording. "Do you think the United States should forbid public speeches against democrary?" (Yes, 54%.) Do you think the United States should allow public speeches against democracy?" (No, 75%).

Another type of error, potentially causing substantive or interpretive difficlties, is the use of response options that each person interprets in his own way. Examples of such response options are the use of words or phrases such as frequently, occasionally, rarely, most of the time, very much, quite a bit, usually, seldom, a great deal, very little, etc.. Presumably words such as always and never mean the same to everyone. But how often is "often"? And how much is "very much"?

Pace and Friedlander, "The meaning of response categories: how often is occasionally, often , and very often?", Research in Higher Education, Vol. 17, No. 3, 1983, addressed this issue using data from the College Student Experiences questionnaire. Participation in various college activities were initially indicated by the responses "never", "occasionally", "often" or "very often". Later in the questionnaire seven of the same activities were responded to as follows: For each of the items below, fill in one of the spaces to the left which best indicates the number of times you have engaged in the activity. These more specific responses were: "never", "once or twice during the year", "about three to six times during the year", "about once or twice a month", "about once a week" and "more than once a week". By this means we were able to show what students meant (number of times) by the more general words. The results, as one would expect, revealed considerable overlap by what was meant by occasionally, often, and very often. But there was also a clear concentration or clustering of responses as one moved from occasional to often, and from often to very often. The meaning or definition of these general descriptors was different, depending upon the topic; but within the same topic the differences between colleges or types of students were quite small. In general, the definition of "occasionally" at one college was similar to its definition at other colleges, given the same topic.

Every respondent knows perfectly well that "very often" is more than "often", and that "often" is more than "occasionally". Thus, the direction of the scale is recognized by everyone. But the specific meaning attached to the labels is an individual judgment. There were few obviously implausible responses -- such as students who initially said "occasionally" or "often" but later said "never"; or students who initially said "occasionally" but later said "more than once a week". These discrepancies constituted from 2% to 10% of the total responses.

Comparative judgments of this sort necessarily reflect some reference group in the mind of the judges. On this questionnaire, we assume that the college peer group is the reference group, and that the answers reflect an awareness of what is customary in one's own behavior and in the behavior of the peer group.

The point of these observations about the subjective definition of response choices is that one should get, if at all possible, some sort of evidence about what people mean by their choices. This same advice applies to opinion polls which ask about degrees of happiness, satisfactions, confidence, or other subjectively defined responses.

PART 2

THREE COLLEGE QUESTIONNAIRES

Efforts to evaluate the influences of college on students' learning and development should draw upon many sources of evidence. For much of this relevant evidence the students themselves are the source; and the most common method for obtaining that evidence is a questionnaire.

Here, for example, are four crucial questions.

1. Who goes? What do we know about the entering students: their high school record and test scores, their family background, financial status, their interests, expectations, aspirations, past achievements, etc.? Some of this information can be obtained from records, but some can be obtained only by asking the students themselves.

2. What do they do after they get there? Some answers can be obtained from college records -- such as, campus residence and major field, but for other sorts of behavior -- such as the time and effort devoted to study, contacts with faculty, involvement in extra-curricular activities, use of the library, etc. -- the answers can only come from students' responses to questionnaires.

3. What's it like? Physical facts -- such as big school, small school, and big city, small town -- are important. So also are students' perceptions of the campus environment or atmosphere. What is stressed? What is expected? How do people relate to one another -- friendly, supportive, or not? Answers to these questions can only come from the students themselves.

4. What do they get out of it? Knowledge, basic skills, and abilities relevant to a career, relevant to personal maturity and life

satisfaction, relevant to civic enlightenment -- these are some of the possible and intended results. Achievement tests, abilty tests, personality tests, etc. can provide some of the answers. It may also be important to find out what the students themselves think they got out of college; and here again one relies on responses to questionnaires.

Questionnaires can, and I think should, be regarded as a form of test or measuring instrument. Many questionnaires are in fact regarded as tests by those who construct them. So, we have tests of attitudes and beliefs, vocational interests, pesonality traits, etc.. A variable or dimension to be measured is defined, sets of items are developed to measure it, and the reliability and validity of the results are determined. The process is similar to the construction of an objective achievement test, or a test of developed abilities such as the Scholastic Aptitude Test. Attitudes, interests, beliefs, etc., are subjective phenomena. The answers one gives to a question about interests or opinions is determined by the individual. The student decides whether he agrees or disagrees with some statement, or likes or dislikes some activity, or person, or condition. The good published tests of personality, interests, or values provide extensive data regarding their reliability and validity -- tests such as the Minnesota Multiphasic, the Omnibus Personality Inventory, Holland's Vocational Preferences Inventory, the Allport-Vernon-Lindzey Study of Values. In some tests of this sort, the authors have included a few items to detect whether a student is giving false or improbable answers -- a practice which recognizes the importance of estimating the credibility of self-reports.

Many of the questionnaires used in studies of higher education are not designed as tests in the classical sense. They consist of sets of items, often grouped or classified under certain topics, but having no underlying or scorable dimension. One finds for example, various items about students use of counseling services, or various items about students opinions of teaching practices, or various items about students attitudes regarding political and economic issues. The items are no doubt regarded as interesting and the answers useful to know. But the content is best described as a classified catalogue rather than as a theoretically or conceptually based set of dimensions or characteristics. The value of the question and the credibility of the answer has to be examined item by item. There is nothing inherently unreliable or invalid about a one-item test. Most public opinion polls are really one-item tests. But it is important to realize that variations in responses are often caused by variations in the phrasing of the question. Slight changes in wording can produce significant changes in responses. Consequently, the meaning of the answers rests on a slender base.

To begin Part 2 we briefly report a tabulation of "missing cases" in three questionnaires for college students. The results illustrate some of the principles and advice given in Part 1, and serve to confirm, with these three current cases, the merit of that advice. Then, the main content of Part 2 is a detailed examination of one questionnaire to illustrate some of the internal and external checks that can be made to assess the reliability and validity of students responses. The content of this one instrument -- Pace's College Student Experiences Questionnaire -- makes meaningful cross checks possible, for it bears upon all four of the topics noted in the introduction to Part 2: Who goes? What do they do after they get there? What's it like? and What do they get out of it?

Missing Cases: What types of questions are not answered?

To provide some current illustrations of non-response to questionnaire items we have examined two widely used instruments, each having the same general purpose and each intended for the same type of population. The first is the Entering Student Survey, distributed by the American College Testing Program. The second is the Student Information Form, distributed by the UCLA Higher Education Research Institute.

Both of these questionnaires are introduced with assurances regarding the confidentiality of the students' responses. The HERI questionnaire says "Identifying information has been requested in order to make subsequent mail followup studies possible. Your response will be held in the strictest professional confidence". The ACT questionnaire says, "The information you supply on this questionnaire will be kept completely confidential. Your name, address, and Social Security number will enable college officials to identify your responses and to contact you directly. The data you supply will be used for research purposes and will not be individually listed on any report. If, however, any question requests information you do not wish to provide, feel free to omit it."

Both questionnaires have many similar and in some cases identical items, for example: age, race, sex, marital status, planned college residence, high school grades, planned college major, planned occupational choice, sources of funding, reasons for going to college. Straightforward identification questions, and questions about specific activities, reasons for going to college, etc. are typically omitted by fewer than 4% of the cases, and often by fewer than 2%. The questions which are omitted by the largest percentages of respondents are ones related to money, religion, expected major field and occupation, and assorted items about personal and social values.

On the HERI questionnaire there are typically about 12% to 13% who do not answer the items about parents income, and sources of funding for college. Many of those items identify specific dollar amounts -- parents total income -- or a specific fact -- listed as a dependent on federal income tax return. No doubt in some instances the students do not know the answers; and perhaps in other instances they regard the question as inappropriate. The ACT does not ask about dollars; it asks whether various sources of funding are a major source, minor source, or not a source. Eleven sources are listed, and about 5 1/2% to 9% of the students do not respond.

The HERI questionnaire asks the students to indicate the religious preference of self, father, and mother. From 15% to 17% do not answer the question.

On the HERI questionnaire 6% of the entering freshmen do not indicate their probable undergraduate major, and nearly 7% do not indicate their probable career occupation. On tne ACT questionnaire the percent of omits is 12% for the probable major and 16% for the probable occupation. The reason for these larger numbers may be owing to the format. The ACT survey has a separate sheet inserted with the questionnaire listing many major fields and occupations. The student finds the 3-digit code that best describes his plans and then fills in these numbers on the questionnaire. Apparently some students just don't bother to do this. On the HERI questionnaire the various fields and occupations are listed on the quesionnaire itself, making the response easier to record. In both cases, however, it seems reasonable to suspect that asking entering freshmen about their probable college major and their probable occupation is not viewed as an answerable question by some students. In fact, on a different part of the HERI questionnaire more than 20% of the students said the chances were very good that they would change their major and change their occupational choice.

In both questionnaires, items about such topics as reasons for going to college and reasons for going to this particular college, were omitted by only 2% to 4% of the respondents in most instances. The ACT questionnaire has a section labeled "college impressions" where students are asked to indicate their agreement with various statements about the college environment -- such as "students at this college are friendly", "this college offers many cultural events and programs". Typically about 3% omit these items; although one wonders about the basis for the answers because often one's impressions, in advance of actual experience, reflect common stereotypes about what college is like.

The HERI questionnaire asks students questions about various political, social, and educational policies -- such as "abortion should be legalized", "college grades should be abolished", "the federal government is not doing enough to control environmental pollution". Typically about 5% to 8% of the students do not answer these questions. Another question asks students to characterize their political views, as far left, liberal, middle-of-the-road, conservative, or far right. About 5% do not answer the question.

For all of the above data, the information about the ACT questionnaire comes from a normative report based on about 16,000 cases in which the number of "blank" responses to every item is listed. For the HERI questionnaire the data come from the 1983 report of freshmen norms in which the data for one sample college are shown, having about 2,300 cases. The complete normative report does not show missing cases.

Except for the questions about major field and probable occupation, the number of "omits" in the ACT questionnaire is generally smaller than in the HERI questionnaire. There may be several factors accounting for this. The ACT questionnaire is shorter. The format and organization are also clearer. Section 1 is labeled Background Information, Section 2 is Educational Plans and Preferences, Section 3 is College Impressions. Although in some parts the print is quite small, each part is enclosed in a box, with the question or topic itself in boldface capital letters. Perhaps more important is the likelihood that most students would not view any of the questions as offensive or intrusive. There is no invasion of privacy of the sort that might influence one to omit the answer or to give a socially desirable answer rather than a more forthright answer. One can easily regard the questions as appropriate to ask of entering students because of the educational relevance of the questions.

The HERI questionnaire, although of the same four-page length as the ACT, has many more items, and the format consequently appears crowded. Also there is no obvious organization or sequence to its questions. The reasons for not answering various questions, however, are probably owing more to the nature of the questions than to the format. Questions about the future -- such as "what is your best guess that you will": graduate with honors, change career choice, transfer to another college, find a job after college in the field for which you were trained, etc.? -- are generally skipped by 5% to 6% of the respondents. Questions about longrange aspirations or values are skipped by 5% to 8% of the students. Also, as noted earlier, questions about political and personal attitudes are typically skipped by 5% to 8% of the students. From one perspective, these are not large percentages; and the conclusions one draws from those who do respond would not be changed in any significant way if everyone had responded. From another perspective, these percentages of missing cases may represent the tip of a deeper and larger problem about the validity of students responses. There is no doubt that some students do not like some of the questions. During the time of student activism in the late 1960s, there were organized student protests against answering the sort of questions that are still included in todays edition. At the end of the questionnaire, 26% of the students do not give permission to include their ID number on any tape for future research or follow-up study. This undermines the validity of the data base for longitudinal studies. Moreover, when one realizes that the response rate to a mailed follow-up questionnaire may be only 50% or even less, then, together with the 26% refusal to be involved, one is left with a respondent population that may be only 1/3 or 1/4 of the population one should have.

Missing cases have also been noted for a third instrument -- Pace's College Student Experience questionnaire. Later in this report a detailed examination of the reliability and validity of responses will be presented. At this point, only the data about missing cases are reported. Most of the questionnaire consists of 142 college activities to which the students respond by indicating whether and how often they have engaged in them during the current school year. These are, for the most part, quite specific events, and apparently quite easy to recall. Based on the responses of about 7,500 undergraduates, the number of missing responses was rarely more than 1%, and never more than 2%. These activities are grouped into scales, usually of 10 items, to which an activity score can be computed. If any item in a scale is not answered no score is computed. The number of missing cases in these scale scores is, in most scales, about 2% or less, and never more than 4%. In other parts of the questionnaire students are asked to indicate how much progress they believe they have made with respect to various goals or objectives, how well satisfied they are with college, and how they would characterize the college environment along various dimensions. The missing cases to these items are often fewer than 1% and never more than 2%. In another brief section of the questionnaire students are asked to indicate about how many textbooks they read, how many non-assigned books, how many essay exams they had, and how many other written reports they made during the current school year. The percent of missing responses was typically from less than 1% to 2%, except among students in not highly selective liberal arts colleges where there were 3% to 4% missing cases. No obvious explanation comes to mind for these somewhat larger percentages. With respect to the usual background items -- age, sex, year in school, etc. -- there are typically no more than 1% or 2% missing cases, except for the questions about the student's major field where the percent of missing cases ranges from 3% to 6% at different types of institutions. Unlike the ACT and HERI questionnaires which are given to beginning freshmen, the CSEQ is answered by undergraduates in general, not just by freshmen, so that most of them do in fact have a definite major field. Why there should be from 3% to 6% omits is a mystery. Of course, not all possible major fields can be listed in the questionnaire so that students may wonder where to classify their own major. Also, especially in the more heterogeneous colleges, and also in the most selective ones, there may be more interdepartmental majors or other special options. Apparently, instead of checking "other" as the proper response, they just omit the item.

The College Student Experiences Questionnaire: A Brief Description.

To understand some of the analyses to be reported next, some knowledge about the content of this questionnaire may be helpful. The questionnaire is meant to be filled out by undergraduates toward the end of the academic year. It is an eight page, 8 1/2 by 11 format, with the cover page indicating what its all about, and stating that "we do not ask you to write your name anywhere in this questionnaire; but we do need to know where the reports come from, and that is why each questionnaire has a number on the back page -- certain blocks of numbers tell us that those questionnaires come from your college". The first 1 1/2 pages consist of "Background Information" -- the usual questions about age, sex, year in school, college residence, major field, parents education; and also time spent on academic work, time on a job, main source of funding for college, grades, race, and citizenship. The next 3 1/2 pages are labeled "College Activities". There are 142 activities, grouped into "scales" or topics labeled library experiences, experiences with faculty, course learning, art-music-theater, student union, athletic and recreation facilities, clubs and organizations, experiences in writing, personal experiences, student acquaintances, science/technology, dormatory or fraternity/sorority, topics of conversation, and information in conversations. The directions are: "In your experience at this college during the current school year, about how often have you done each of the following?" The responses are "never", "occasionally", "often", and "very often". The activities are fairly specific so that the student would presumably recall accurately whether he had ever done them; but of course the frequency estimate is entirely a subjective response. Examples of activities are as follows: read something in the reserve book room or reference section, made an appointment to meet with a faculty member in his/her office, summarized major points and information in readings or notes, gone to an art gallery or art exhibit on the campus, meet your friends at the student union or student center, played on an intramural team, worked on a committee, asked other people to read something you wrote to see if it was clear to them, sought out a friend to help you with a personal problem, made friends with students from another country, practiced to improve your skill in using some laboratory equipment, gone out with other students for late night snacks, talked about current events in the news, referred to something a professor said about the (conversation) topic.

The next brief part of the questionnaire asks students to report how much reading and writing they have done, and how well satisfied they are with college.

The next main topic is labeled "The College Environment". This consists of eight rating scales on which students report their impressions of the emphasis or stress there is in the environment on such aspects of students' development as academic and scholarly qualities, esthetic and creative qualities, being critical and analytical, vocational and occupational competence, and the general relevance and practical values of the courses; also their impressions of the personal relationships in the environment, ranging from supportive, helpful, considerate to alienated, unsympathetic, and rigid with respect to the relationships among students, between students and faculty, and with administrative personnel. Finally, the last section, labeled "Estimate of Gains", lists 21 goals or objectives of college education and asks students as follows: "In thinking over your experiences in college up to now, to what extent do you feel you have gained or made progress in each of the following respects?" The responses are "very little", "some", "quite a bit", and "very much". Here again, the responses are entirely subjective.

From one perspective, this College Student Experiences questionnaire includes features that some think should be avoided, if possible. The ratings are entirely subjective, the estimates of the amount of gain are entirely subjective, and the reports of frequency of participation in activities are entirely subjective. What follows next are examples of how subjective responses can be objectively validated.

Test-Retest Comparisons

In the absence of any major changes in the campus environment or facilities or admissions policy, one would expect some consistency in the amount, scope, and quality of effort revealed by students' responses to the activities scales by different but comparable samples. Recently, several colleges have used the College Student Experiences questionnaire twice -- once in 1984, and again in 1985. Such comparisons are not, strictly speaking, an indication of the reliability of self-reports. The answers come from different students and from a different time. Some changes in the responses may reflect true changes, not random changes or errors of measurement. Nevertheless, if one found substantial variations in the responses of two different but similarly selected samples, even though a year apart, one would worry about the dependability of the results.

The best test-retest example comes from Denver University. It is best in the sense that the sample size was fairly large -- 635 in the spring of 1984 and 661 in the spring of 1985. The samples were selected in the same way, the response rate was similar, and the two goups did not differ in such population descriptors as age, sex, year in school, major field, grades, residence, transfers, etc.. No attempt is made to compare the responses to every item in the questionnaire. Rather, to get a general indication of stability, comparisons are made between the mean scores on each of the 14 activity scales, and the mean ratings on each of the environmental characteristics. It is not appropriate to report the scores on these matters, but it is permissible to report the differences between the 1984 scores and the 1985 scores. The second test-retest example comes from Case Western Reserve University -- with a sample of 779 students in the spring of 1984 and of 376 in the spring of 1985. The characteristics of the two samples are nearly identical with respect to age, sex, year in school, transfer status, major field, etc.. The third example comes from Keuka, a small college for women in upstate New York -- with 148 in the

1984 sample and 130 in the 1985 sample. The groups were similar in all respects except one: the 1985 sample had a larger proportion of freshmen.

On the 10-item activity scales the possible range of scores is 30 points; 36 points on the three 12-item scales; and 20 points on the one 6-item scale. The typical standard deviations are 5.7 on the 10-item scales, 6.0 on the 12-item scales, and 3.2 on the 6-item scale. A glance at the list of differences in the table quickly reveals that at all three schools the magnitude of differences is usually less than one point. This is true of 13 out of 14 scales at Denver, all 14 at Case Western Reserve, and 10 of the 14 scales at Keuka. In fact, at Denver the difference in mean scores between the 1984 and 1985 samples is .5 or less on 10 of the scales, at Case Western Reserve the differences are .5 or less on 13 of the 14 scales; and at Keuka on 6 of the 14 scales.

If comparable scores from comparable samples, even though a year apart, is an indication of test reliability, then obviously these student self-reports are very stable and dependable. At Denver, where there is a significant difference of 2.4 points on the Student Union scale, the explanation is a good example of a change in results owing to a change in conditions. During 1984 at Denver a new student union and activity center was under construction; 1985 was the first full year of its operation, and, not suprisingly, the activity score for students' use of the union increased significantly. At Keuka the differences between mean scores, although greater than 1.00 on four of the scales, are not statistically significant.

From these comparisons, one can surely conclude that self-reported activities and self-reported ratings of environmental characteristics are dependable and consistent.

Test-Retest Mean Differences -- 1984 and 1985
In Activity Scale Scores and Environment Ratings

Activity Scales	Denver Univ.	Case Western Reserve	Keuka College
Library Experiences	.3	.4	.7
Experiences with Faculty	.3	.1	1.3
Course Learning	.6	.2	0
Art, Music, Theater	.5	.4	.1
Student Union	2.4	.2	.4
Clubs and Organizations	.5	.9	.5
Athletic and Recreation	.1	.4	0
Experience in Writing	.5	.2	.2
Personal Experiences	0	.3	.7
Student Acquaintances	.6	.1	1.4
Science/Technology	0	.5	1.2
Dormatory or Fraternity/Sorority	.9	0	1.5
Topics of Conversation	.5	.4	.6
Information in Conversations	.1	.2	.8
Environment ratings			
Academic	.1	0	.1
Esthetic	.1	.2	.2
Critical/analytical	.1	.1	.1
Vocational	.2	0	.2
Personal Relevance	.2	.2	.1
Students	.3	0	.1
Faculty	1.3	0	.2
Administration	.1	0	.4

External Validity: Self-reported gains vs objectively known achievement

Over the past 50 years hundreds of thousands of college students have taken objective achievement tests in various college subjects, tests constructed by national testing agencies. Certain conclusions from all this testing are so consistent, and so obvious, that it almost seems unnecessary to state them; but if one is to document that self-reported achievement corresponds to objectively tested achievement, then some examples of the test evidence must be given. The examples that follow are reported in Pace, *Measuring Outcomes of College*, Jossey-Bass, 1979.

The first example shows the relationship between credit hours and test scores from the Pennsylvania study in 1928. The obvious conclusion is that students learn what they study, and the more they study the more they learn. Students who had the most credit hours in the natural sciences had the highest test scores on the natural sciences test items. The same is true for credits and scores in language, literature and fine arts, and also for credits and scores in social studies.

Credit Hours and Test Scores: 4500 Seniors from 45 Colleges in Pennsylvania, Tested in 1928

Natural Sciences Credits	Natural Sciences Test Scores
High: 55 or more	120
Statewide average: 37	78
Low: 6 or fewer	46

Language, Literature, and Fine Arts Credits	Language, Literature or Fine Arts Test Scores
High: 67 or more	250
Statewide average: 42	168
Low: 12 or fewer	111

Social Studies Credit Hours	Social Studies Test Scores
High: 97 or more	292
Statewide average: 52	241
Low: 12 or fewer	196

The second example, some forty years later, comes from the Area tests of ETS's Undergraduate Assessment Program. The test results are based on 47,000 seniors from 211 colleges in the years 1969-1971. For each of the three Area tests -- Humanities, Natural Sciences, and Social Sciences -- the average score for all seniors is compared with the average score of seniors whose "area of interest" corresponds to the subject matter of the test. The scores are standardized scores in which the standard deviation is 100 points. In the humanities area the differences between the two groups is 55 points. In the natural sciences area the differences are 57 points and 66 points. In the social sciences area the difference is 2 points. The sub-group is also part of the total group; and since 60% of the total group of seniors are also in the social science interest group, the difference is necessarily small.

UAP Area Tests: Approximately 47,000 Seniors from 211 Colleges in the Years 1969-1971

	Humanities Scores
All seniors	470
Seniors whose area of interest is in humanities (21% of all seniors)	525

	Natural Sciences Test Scores
All seniors	480
Seniors whose area of interest is in biological sciences (12% of all seniors)	537
Seniors whose area of interest is in physical sciences (7% of all seniors)	556

	Social Sciences Test Scores
All seniors	446
Seniors whose area of interest is in social sciences (60% of all seniors)	448

The third example comes from the UAP tests in designated major fields rather than from the more general Area tests. These results are shown in relation to the number of courses students had taken in their major -- fewer than eight vs eight or more courses. It is unlikely that , in one's major field, one would have fewer than six courses and still qualify as a major. Most likely, the comparisons are between students who have had 6 or 7 courses vs those who have had 8 to 12 courses. Again, the more one studies a subject the more one knows about it.

Given these obvious conclusions from decades of achievement testing, one can surely use them as external validity in relation to self-reported achievement. The College Student Experiences questionnaire, in the section labeled Estimate of Gains, lists 21 objectives. Students are asked "to what extent do you feel you have gained or made progress in each ...?" They could check "very little", "some", "quite a bit", or "very much". Not all of the objectives are associated with a specific major field, or even with any course-related experience -- objectives such as "ability to fuction as a team member", "ability to learn on your own, pursue ideas, and find information you need". There are, however, eight goals that are related to the curriculum, and specifically to a major field within the curriculum, or to a specific type of subject-matter. These subject-matter goals include Fine Arts, Literature, English (writing), Science, Technology, Computers, Quantitative thinking, and Philosophy/Cultures. If student self-reports are valid they should show the same results that test scores show -- higher achievement (progress) by students whose major field is similar to the objectives as compared with the average of all students -- and this is exactly what the results do, in fact, very clearly show.

Scale Scores of Seniors on Major Field Tests of the Undergraduate Assessment-Program, 1969-1971, Related to Number of Courses Taken in the Major Field

	Fewer than eight courses	Eight or more courses	Difference
Sciences and Engineering Tests			
Biology	539	566	+ 27
Chemistry	510	539	+ 29
Engineering	506	528	+ 22
Humanities Tests			
History	468	491	+ 23
Literature	455	491	+ 36
Philosophy	514	551	+ 37
French	448	486	+ 38

Note: The number of students tested varies by major field, ranging from approximately 1,000 to 8,000.

The data presented here are from a composite of 13,650 undergraduates from 49 colleges and universities who responded to the CSEQ in the spring of 1983, 1984 or 1985. Only those colleges that had given the questionnaire to all four undergraduate classes are included in these composite results. Note also that knowledge or progress is necessarily less among freshmen or sophmores who have not yet accumulated many credits in what is or will be their major field, than it would be among juniors and seniors who, by definition, have accumulated a much larger number of credits in their chosen major field. For some, then, the "major" may reflect an "area of interest" and for others it may be a course program nearly completed.

In the list below, the first four goals are related to the subject matter of arts and humanities, and the second four goals are related to the sciences. Among students who identified their major field as "Arts (art music, theater, etc.)", 92% reported substantial gain ("quite a bit" plus "very much") toward the objective "developing an understanding and enjoyment of art, music and drama". This high percentage contrasts with 29% among students in general. For the objective related to literature, 74% of humanities majors reported substantial gain compared with 33% for students in general. With respect to writing clearly and effectively, 80% of the humanities majors reported substantial progress compared with 57% of students in general. The goal described as "becoming aware of different philosophies, cultures, and ways of life" is not so clearly tied to classroom subject matter in the sense that students' interpersonal experiences might well contribute significantly toward its attainment; but presumably courses in philosophy, history, anthropology, etc. would also be influential. The results show substantial progress reported by 70% of humanities majors, and 64% of social sciences majors, compared with 51% by students in general.

Comparisons of Self-Reported Gains with Known Data About Achievement

Gains in developing an understanding and enjoyment of art, music, and drama

ARTS majors reporting substantial gains	92%
average of all students	29%

Gains in broadening your acquaintance and enjoyment of literature

HUMANITIES majors reporting substantial gains	74%
average of all students	38%

Gains in writing clearly and effectively

HUMANITIES majors reporting substantial gians	80%
average of all students	57%

Gains in becoming aware of different philosophies and cultures

HUMANITIES majors reporting substantial gains	70%
SOCIAL SCIENCE majors reporting substantial gains	64%
averge of all students	51%

Gains in understanding the nature of science and experimentation

BIOLOGICAL SCIENCES majors reporting substantial gains	85%
PHYSICAL SCIENCES majors reporting substantial gains	76%
average of all students	36%

Gains in understanding new scientific and technical developments

BIOLOGICAL SCIENCES majors reporting substantial gains	74%
PHYSICAL SCIENCES majors reporting substantial gains	66%
ENGINEERING majors reporting substantial gains	66%
average of all students	31%

Gains in acquiring familiarity with the use of computers

COMPUTER SCIENCE majors reporting substantial gains	90%
ENGINEERING majors reporting substantial gains	65%
average of all students	32%

Gains in quantitative thinking -- understanding probabilities, proportions, etc.

PHYSICAL SCIENCES majors reporting substantial gains	68%
ENGINEERING majors reporting substantial gains	68%
average of all students	47%

In "understanding the nature of science and experimentation", substantial progress is claimed by 85% of biological sciences majors and 76% of physical sciences majors, compared with 36% for students in general. A similar result is shown for "understanding new scientific and technical developments", with percentages of 74% and 66% for scientific and technical majors, compared with 31% for the average of all students. The contrasting percentages for the goal "acquiring familiarity with the use of computers" are even sharper -- 90% of majors in computer science indicating substantial progress compared with 32% for the average of all students. With respect to quantitative thinking, students majoring in fields where much quantitative thinking is required -- engineering, and physical sciences -- are most likely to claim substantial progress (68%) compared with 47% among students in general.

All of the above results document the external validity of students self-reports. When asked to rate their progress toward goals that are obviously related to the subject matter of college courses, the ratings are totally congruent with what we know from achievement test scores and from the relationship between credit hours or amount of study and measued achievement.

One does not know the actual level of measured achievement (standardized test scores) that is associated with the students' self-estimate of gain. No doubt some students who rate their own progress as "quite a bit" may have higher achievement test scores than students at another college who rate their progress as "very much". Such discrepancies probably reflect institutional differences in academic selectivity and academic demands. The same variability applies to credit hours vs test scores. While it is true that the more courses one takes in a subject the more one is likely to know about it, it is also true that some students who have taken 5 or 6 courses may get higher test scores than some students who have taken 9 or 10 courses. But the averages are consistent. Sorting students according to course work (major field) or according to achievement test scores (major field) or according to self-reported progress (in major fields) all produce the same conclusions.

Internal Reliability: Consistency in responses to similar items

In the Science/Technology activity scale there are three activities that clearly involve conversation about science. These items, together with the percent of students who said they engaged in them frequently, are shown below:

	% Frequently among			
Science activities	Bio.Sci. majors	Phys.Sci. majors	Engr. majors	Average of all students
Tested your understanding of some scientific principle by seeing if you could explain it to another student.	70	69	69	34
Showed a classmate how to use a piece of scientific equipment	43	35	34	18
Attempted to explain an experimental procedure to student	43	42	41	15
Conversation topic				
Science -- theories, experiments, methods	57	53	58	21

The conversation item appears in a different part of the questionnaire. Presumably, the percent of students who say they have frequently talked about science with other students should have some similarity to the percent who said they had tried to explain a principle, a procedure, and the use of equipment to another student. The responses were, in fact, very similar.

A similar comparison can be made in the arts. In the activity scale labeled Art, Music, Theater there are three "talk about" items, and later, among the conversation topics there is a topic described as "Fine arts - painting, theatrical productions, ballet, symphony, etc.". Here are the results.

Art, Music, Theater activities	% Frequently among Arts majors	Among all students
Talked about art (painting, sculpture, architecture, artists, etc.) with other students at the college	68	17
Talked about music (classical, popular, musicians, etc.) with other students at the college	73	35
Talked about the theater (plays, musicals, dance, etc.) with other students at the college	58	20
Conversation topic		
Fine arts -- painting, theatrical productions, ballet, symphony, etc.	78	17

Similar but not identical questions produce similar but not identical answers. The general congruence shown above can be regarded as an indication of internal reliability.

Internal Validity: finding plausible connections

For the attainment of many goals of higher education there is no readily available objective documentation and in some cases no external evidence at all. One can use tests and credits when the goals are related to the curriculum or to particular courses and major fields. But what does one use for an external criterion when the goals are self-understanding, understanding others, good health habits, functioning as a team member, etc.?

In this part of the report several examples of internal consistencies that should be found are used to make the case for the credibility of self-reports. The first example is surely a connection that should exist. The activities -- setting performance goals, following a regular exercise schedule, and keeping a record of progress -- are, to a considerable extent, behavioral indicators of what is involved in "developing good health habits and physical fitness". The tabulations show that students who report "very much" progress toward this goal are much more likely to set goals, follow a schedule, and keep a record than students whose self-rated progress is lower.

Similar tabulations are shown for several other goals. In every case, the behavior that surely should contribute to students' estimated progress is clearly related to that progress. The differences in percents between "very much" and "very little" are uniformly large, the one being from 2 to more than 6 times larger than the other.

If student responses to the gains items or to the activity items were capricious or unreliable or invalid, the congruent and plausible connections shown in the tables below would not occur. If what should be true is also true empirically, the credibility of self-reports is further documented.

Goal: Developing good health habits and physical fitness

Activity	Very Much	Quite a bit	Some	Very Little	Average of all students
	Percent engaging in activity frequently among students who rate their progress as:				
Set goals for your performance in some skill (athletic)	77	58	36	23	45
Followed a regular schedule of exercise, or practice in some sport, on campus	71	53	28	14	38
Kept a chart or record of your progress in some skill or athletic activity.	28	15	6	3	11

Goal: Ability to function as a team member

Activity	Very Much	Quite a bit	Some	Very Little	Average of all students
	Percent engaging in activity frequently among students who rate their progress as:				
Used outdoor recreational spaces for casual and informal group sports	40	27	15	7	23
Used facilities in the gym for playing sports that require more than one person	42	30	18	10	26
Played on an intramural team	36	26	15	7	22

Goal: Understanding yourself -- your abilities, interests, and personality

Activity	Very Much	Quite a bit	Some	Very Little	Average of all students
	Percent engaging in activity frequently among students who rate their progress as:				
Read articles or books about personal adjustment and personality development	38	25	20	15	28
Asked a friend to tell you what he/she really thought about you	33	21	14	12	23
Identified with a character in a book or movie and wondered what you might have done under similar circumstances	56	44	36	32	46

Goal: Understanding other people and the abiltiy to get along with different kinds of people

Activity	Very Much	Quite a bit	Some	Very Little	Average of all students
	Percent engaging in activity frequently among students who rate their progress as:				
Made friends with students whose interests were very different from yours	73	57	38	32	59
Made friends with students whose family background (economic and social) was very different from yours	78	63	44	36	63
Had serious discussions with students whose political opinions were very different from yours	45	33	26	22	35

Goal: Becoming aware of different philosophies, cultues, and ways of life

Activity	Very Much	Quite a bit	Some	Very Little	Average of all students
	Percent engaging in activity frequently among students who rate their progress as:				
Made friends with students whose race was different from yours	62	50	40	33	46
Made friends with students from another country	50	24	24	20	31
Had serious discussions with students whose philosophy of life or personal values were very different from yours	64	48	33	25	43
Had serious discussions with students whose religious beliefs were very diffrent from yours	55	40	28	22	36
Had serious discussions with students from a country different from yours	42	25	15	13	23

Summation

Claims for the credibility of student self-reports can be supported by:

1. Evidence of test-retest consistency.
2. Congruence with externally known facts, when such facts are available.
3. Similar answers to questionnaire items, when questions are asked in more than one way.
4. Congruent, or expected, connections between items that presumably should have connected responses -- as for example between behavior and progress.

One final note may be important. Some psychometricians and survey research analysts point out that the context within which questions are asked may influence the response. In the College Student Experiences questionnaire, some people might claim that the answer to the Estimates of Gains items might be "contaminated" by all the preceding items. The gains might be reported differently if they were asked separately, or without the prior context in the questionnaire. There is, however, a very different way of regarding this matter. If the gains items were presented alone, without any context, the responses would be all the more influenced by personal idiosyncrasies, and hence all the more likely to produce random variations. By putting the gains items at the end of the questionnaire, one increases the credibility of answers. Everyone comes to these items with the same background, having recalled one's behavior during the year, having characterized the college environment, having reported how much one has studied, what grades one has received, etc. so that, for everyone, the estimate of gains becomes a more or less commonly based and thoughtful summary of the college experience, and therefore has a greater reliability.

Finally, as a capstone illustration of what can be done to assess the reliability of self-reports, we apply some multivariate statistical procedures which bear upon the predictive and construct validity of certain parts of the College Student Experiences questionnaire.

Multivariate statistical procedures

In this part of the report we describe the use of common multivariate statistical procedures to assess the validity of self report. The goal is to demonstrate that for surveys that allow internal validity checks, one can go beyond item-by-item validity to assessing the validity of self report at the construct level. These techniques are applied to a sample of 6,000 undergraduates who provided responses to the College Student Experiences Questionnaire (CSEQ).

Three techniques were applied to two types of scales and one background variable of the CSEQ. The background variable is academic major coded as: 1) Arts; 2) Biological Sciences; 3) Business; 4) Computer Science; 5) Education; 6) Engineering; 7) Health related fields; 8) Humanities; 9) Physical Sciences; 10) Social Sciences. The two types of scales are composed of 13 subscales from the Quality of Effort (QE) measures, and 21 items from the Estimate of Gains (EG) measures.

The first statistical procedure is discriminant analysis with special attention paid to the classification phase of the analysis. The discriminating variables are the EG items while the classification variable is academic major. Since the number of undergraduates in each major are not the same, special a priori weighting is given to the samples during the classification phase. The rationale for using discriminant analysis in this context is that those who major in certain academic disciplines probably make the most gains in those areas related to that discipline. Hence, if one knows a student's set of responses to the gains items, one should be able to predict that individual's major. To the extent that this is true, it might be argued that the EG measures provide valid self report measures of gains.

The second procedure is canonical correlation analysis applied to the QE subscales and the EG items. This procedure attempts to find a set of linear combinations (canonical variates) within a scale that are maximally correlated with linear combinations formed from the other scale. To the extent that these canonical variates are interpretable, we would expect high canonical correlations among those variates from each set that have something in common. Often it is the case that canonical analysis obscures the simple factor structure that might exist within a set of items. To address this problem, the third procedure is to factor analyze the QE subscales and EG items separately, rotate the factors for maximum interpretabilty, calculate factor scores, and correlate factor scores using simple Pearson correlations. It is expected that Pearson correlations should show high correlations among those factors that are substantively related.

Discriminant and classification analysis were performed using ten academic majors and twenty EG items. The EG items were chosen to correspond as closely as possible to the academic majors, hence the item related to gains in vocational training was omitted since no major was uniquely vocational.

The results of the classification phase are displayed in the following table. The table shows the percentage of those who were classified into their known majors on the basis of the discriminant analysis. The diagnonal represents the percentage of correct classifications, while the off-diagonal represents the misclassifications. It can be seen that the EG responses tend to do well in predicting academic major. For example, 61% of all art majors were correctly classified as being art majors on the basis of the discriminant analysis. Certain incorrect classifications did occur; but the misclassifications were in a sensible direction. For example, physical science majors (including chemistry and math) were more often classified as biological science majors (including biochemistry) or engineering majors. Overall, these results lend support to the claim that self report of gains as measured by the EG data are valid in that they adequately predict a relatively objective measure of academic field where most gains should occur.

The results of the canonical analysis are displayed in the next table. Here, only the first two canonical variates extracted from each set of measures are presented. Note that the standardized canonical coefficients can be loosely interpreted as factor loadings.

Inspection of the standardized canonical coefficients for the QE subscales suggests that the first canonical variate is dominated by the QE subscale related to Science and Technology. The first canonical variate to the EG data appears to be dominated by those items related to computer knowlege and Science/Technology. The squared canonical correlation between

Classification Analysis of Academic Major on Basis of Discriminant Analysis*

	Predicted Major										
True Major	Art	Bio	Bus	C/S	Ed	Engr	Heal	Hum	PhyS	SocS	Total
Arts	61	4	11	1	0	0	0	11	0	11	100
Bio	1	52	7	1	0	15	13	0	0	10	100
Bus	2	1	73	4	2	3	2	1	0	12	100
CompSci	1	1	33	45	0	9	0	1	0	1	100
Educ	9	2	34	2	15	3	9	7	0	18	100
Engr	1	10	14	8	0	57	4	0	1	5	100
Health	2	20	15	0	4	4	34	1	0	19	100
Human	8	3	11	1	3	1	3	38	0	32	100
Phs/Sci	0	33	17	7	1	25	5	1	3	7	100
Soc/Sci	4	7	28	2	4	3	4	9	0	39	100
TOTAL %	5	11	33	6	3	11	7	6	0	17	100

* Entries are in percentages.

Standard Canonical Variates for
QE Scale and EG Scale Items

EQ Canonical Variates

Subscales	QE1	QE2
Library	.02	.05
Faculty	- .06	.10
Course Work	- .05	.15
Art, Music Drama	- .16	.66
Student Union	.00	- .05
Recreation	- .00	- .05
Clubs	.00	.04
Writing	- .14	- .03
Personal Experiences	- .14	- .03
Acquaintances	- .04	.01
Sci/Tech.	.94	.21
Conv. Topics	.03	.21
Information	- .06	.20

EG Canonical Variates

Items	EG1	EG2
Professional Sci or Scholarly	.03	.10
General Education	- .02	- .02
Career Development	- .07	.03
Art, Music, Drama	- .14	.64
Literature	- .07	.16
Writing	- .17	- .13
Computers	.51	.09
Philosophies/ Culture	- .08	.09
Ethical Standards	- .02	.08
Personality	- .03	.01
Understanding People	- .05	.08
Team Work	.01	.04
Physical Fitness	- .03	- .06
Science Experim.	.27	.01
Science/Technology	.32	.15
Technology/Hazards	.00	.14
Analytical Thinking	.05	- .01
Quantitative Thinking	.16	- .15
Similarities and Differences	- .11	.20

these two canonical variates is statistically significant [$R^2 = 0.61$, $F (260, 44745) = 42.979$, $p < .000$].

Inspection of the second canonical variates for both sets of measures reveals a similar consistent picture. The second canonical variate for the QE scale is related to art, music, and theater, while the second canonical variate of the EG responses is related to gains in understanding art, music, and drama. Again, the squared canonical correlation between this pair of variates is statistically significant [$R^2 = 0.38$, $F (228, 41681) = 29.011$, $p < .000$]. Subsequent canonical variates were difficult to interpret.

It can be seen that the canonical analysis gives a useful, though limited, picture of the internal validity of the two self report measures. Again, it should be noted that this procedure examined validity of self report at the construct level, where the canonical variates can be taken as representing the constructs, though perhaps not in the factor analytic sense.

On the basis of previous research, four factors of the QE scale and five factors of the EG items were independently extracted and obliquely rotated to simple structure. The four factors of the QE scale were labeled 1) Personal/Social; 2) Academic/Intellectual; 3) Clubs/Organizations; 4) Science. The five EG factors were labeled 1) Personal/Social; 2) Science/Technology; 3) General Education; 4) Intellectual; 5) Vocational. A matrix of Pearson correlations among the factor scores obtained from the factor analysis is displayed in the next table. Although most of the correlations are large and significant, those that are highest are among

Intercorrelations Among Factor Scores

		Quality of Effort Factors			
		P/S	A/I	C/O	Sci
Estimated Gains Factors	P/S	.50	.42	.44	.11
	S/T	.19	.13	.04	.62
	G/E	.42	.45	.31	.07
	Intel	.36	.36	.22	.43
	Voc	.29	.29	.24	.25

those factors that have something in common. For example, the gains in personal and social development factor is most highly correlated with QE factor measuring personal and social aspects such as student acquaintences, personal experiences, and topics of conversation.

Two points can be made with regard to the above analyses. First, the application of multivariate statistical procedures for assessing broad construct validity of self report has potential. It should be pointed out however, that construct validty in the factor analytic sense was only explained via the factor score correlations. Secondly, with respect to the CSEQ, and the QE scales and EG items in particular, evidence does exist for claiming a certain degree of validity in these self report measures. The result of all three analyses present a picture of a questionaire that is consistent with respect to self report predictive validity and self report construct validity.

CONCLUDING COMMENTS

This report is obviously not a definitive document about the credibility of questionnaire survey responses by college students. It has aimed, nevertheless, to show that there are many ways to confirm the accuracy, reliability, and validity of student self-reports. It has also noted, from examples in higher education, and from examples in the larger area of public opinion polls and other general surveys, some of the common sources of measurement errors and errors of substance.

In academic surveys the high proportion of students who do not reply to the questionnaires they have received is a most serious problem. One wonders whether rigorous follow-up messages would make a big difference, or whether the magnitude of the non-respondent problem reflects a deeper rejection of such inquiries. Twenty years ago one could expect about two thirds of college students to respond to a questionnaire. Today, one is grateful if 50% respond. Times change. Nearly 50 years ago, in a study I directed of former university students, including some who had graduated and some who had not, we got returns from 70% of those who received the questionnaire. The questionnaire was 52 pages long and took about two hours to answer. But that was before the invention of television! (Pace, They Went to College, University of Minesota Press, 1941).

My own belief is that the likelihood of good returns is enhanced by the recipients' opinions about the importance of the topic, its perceived relevance to one's experience, one's regard for the source of the inquiry and the likely use or value of the results, the clarity of questions and the ease or confidence one has in answering them, and the overall attractiveness of the design, format, typography, etc. of the instrument. I also believe that unless these conditions are reasonably well met, even vigorous follow-up efforts will have little influence on the response rate, and even when some increase in response rate is achieved I would be skeptical about the integrity of those added responses.

Perhaps the second most common weakness in questionnaires by academic organizations is the inclusion of questions that are quite likely to have unreliable or invalid answers. These may be questions about vague concepts, questions about topics that students have not previously thought about, questions about values or life goals or future plans. Similar weaknesses are evident in public opinion polls that ask for opinions about ambiguous or undefined concepts such as national defense, foreign aid, national health, etc. The unfortunate consequence is that pollsters and public alike think that the results reflect public attitudes toward the matter, when in fact the topic is complex, can be phrased in a variety of proper ways, and all one has done is to tally answers to the particular question which is not well or uniformly interpreted in the first place. Questions about future expectations can be very clear -- for example, "Do you expect to have any (more) children?" But it is difficult to know just what is being measured or revealed by answers to questions that different people can interpret in different ways.

A final issue is the use of single questions versus the use of scales or combinations of questions that can be added together to produce a score or index. Commercial agencies rely on single items. Scale development is complex, time-consuming, and costly; and for public opinion polling agencies the presumed benefit is not worth the price. A scale is not always better (more reliable and valid) than a single item. In most academic surveys, however, the topics of inquiry tend to be rather global rather than narrowly explicit. In these cases there is merit in thinking about questionnaire construction in ways somewhat similar to thinking about test construction.

Whatever the topic of inquiry, it may well be that one of the most important elements to consider in writing the questions is the nature of judgment required to answer them. If the judgment or thought process is one of recall, is the thing or condition to be recalled clear and are the respondents able to recall accurately? If the judgment is one of comparison, is the base for the comparison clear and do the respondents have the experience or knowledge needed to make the comparison with reasonable confidence. If the judgment or thought process to answer the questions is one of generalizing or inferring, do the respondents understand what is to be generalized? Many survey questions would probably yield better answers if the writers always asked themselves such questions as: Does the respondent have the knowledge or experience to give a useful answer? Will different people interpret the quesiton in the same way? Will the answer be accurate? What can I conclude or interpret from the answers to this question?

The quality of questionnaire answers (reliability, validity, credibility) depends most of all on the quality of the questions.

ASSESSING INSTRUCTIONAL OUTCOMES

by

Eva L. Baker, Center for the Study of Evaluation
University of California, Los Angeles

and

Harold F. O'Neil, Jr.,
University of Southern California

ASSESSING INSTRUCTIONAL OUTCOMES

Eva L. Baker, Center for the Study of Evaluation,
University of California at Los Angeles

and

Harold F. O'Neil, Jr., University of Southern California

This chapter is addressed to the topic of assessing instructional outcomes. It occupies, conceptually, an interesting point in the consideration of instructional technology. On the one hand, the hallmark of technology is its repeatable utility based upon its use of verified knowledge produced from research. Assessment is clearly a requirement to determine if one has a technology that works. On the other hand, in practice, the serious consideration of assessing educational outcomes is often overlooked in the excitement of exploring innovation or in the day-to-day tedium of producing sufficient amounts of courseware or other instructional products on schedule and within budgetary constraints. Because of the lack of attention to the issue of educational outcome assessment, measuring outcomes in the recent history of instructional design has been treated routinely, more as an historical obligation than as a tool integrally related to the improvement of instructional effectiveness. For this reason, it is important to see that the measurement of instructional outcomes has two critical functions: 1) it is both a means to assess how well the product, courseware, or other technology performs, and 2) it is a mechanism to intervene in and to improve the process of instructional design and development itself.

Basic to the understanding of the assessment of instructional outcomes is the role of tests. Unfortunately, the term tests conjures up some of the least useful forms of assessment and restricts the instructional designer's view of the full range of information useful for making important inferences about the effects of learning. While basic understanding of testing is important, and will be treated in this chapter as well, it is not sufficient. It is more important to think broadly first about information needed to make instructional decisions, and secondly, about the inferences one can draw from such information to make decisions about the quality of instructional efforts.

At the heart of both the information base and the inferencing process is the notion of validity, and it should be the overriding concern in the process of assessment. The formats of assessment, where they take on the coloration of typical tests or even look very different from the tests we have seen and taken in school, are at best secondary concerns. Our intent is to raise the salience of assessment in the entire design and development process by identifying the critical attributes of valid information and inferences. Then we will move to discussion of the various sorts of testing and other assessment options and consider their strengths and limitations against a framework of validity.

Measurement: The Basics

Without deeply investigating the psychometric theory underlying measurement, an instructional designer can still treat the assessment issue seriously. A few straightforward points need to be reviewed. First, all measurement is imprecise. Everything we infer is exactly that: inferencing about learning that has occurred (or is potential) in the learner. As measurement begins to use some of the newer techniques in the biotechnical area, readings of magnetic fields, heat, and other electrical brain activity, then we may appear to be closer to direct measurement of learning. But since we are dealing with the mind, we will still remain in the land of inference and inevitably be left to piece together what has actually been experienced by the learner.

Second, a good deal of what is measured is inaccurate because we have chosen the wrong thing to measure. We may have chosen an approach inappropriate to the subject matter, chosen to measure performance in a particular way because of its practicality and convenience rather than for reasons related to the accuracy of assessment. So even if we were to improve our precision, we would err by selecting, some of the time, the wrong matter to which to attend.

Third, we must remember we are dealing with people, not plastics. People are dynamic; all change from second to second. The meanings they ascribe to events become successively refined and restructured with experience. They are blurry targets for precise metrics. As we all know, people not only change continuously but they differ from one another enormously. They have color preferences, various language facilities, and predispositions to certain subject matter content, for instance. They also have very different perceptions of themselves as learners, of their abilities to succeed, and of the reasons they succeed and fail (see Weiner, for example, 1979). Some are desperately anxious when they are given tests (O'Neil & Richardson, 1980), some worry about only one sort of test, like essays or multiple choice, and some are relatively accepting of whatever tasks come their way.

People also think in different ways. Their approaches differ not only as a function of the level of ignorance or expertise they have about a single subject, but their general background or world knowledge. They also approach problems very differently. One style is methodical and analytic; learners of this sort see the world in terms of components that get built up or decomposed into smaller parts. Other see the world in broad patterns, seek integration, use metaphors, and focus on the whole rather than its parts. And many people use both approaches described, switching within the same problem sometimes to understand through one or another means. These approaches were described simply and archetypally to make a point. But, it should be remembered that a good deal of style of learning comes automatically to the learner. Only infrequently is learning style a volitional matter, although there have been moderately successful attempts to affect the use of various learning strategies (O'Neil, 1979; Danseresu et al, In Press; Moore et al, 1985).

Our primitive measurement tools will miss a good deal of this complexity. So even if we had precise methods, and were confident that we were assessing the correct type of learning, we would still be sure to miss a good deal of the truths of what our effects have been.

It is for all these reasons that we can not claim to have proved that our instruction is effective, just as we cannot prove that a scientific theory is right. We have to repeat our measurements, find multiple approaches to assess the outcomes we are intending, and still couch our conclusions tentatively. In the educational marketplace, of course, tentativeness goes by the board. Instructional designers compete with claims about materials proven effective, quality assurance and other slogans designed to loosen resources from program managers either in business or in government. But, in the secret recesses of one's own mind, it is important to know what we don't know, even if our roles or organizations require different public proclamations.

Purposes of Assessment

Central to the problem of assessing instructional outcomes is the issue of purpose: for what purpose are we to assess outcomes? One common enough response is to assess the quality of our intervention in meeting its particular goals. If a program or system is devoted to teaching reading comprehension, then it is appropriate to assess the extent to which reading comprehension ability is affected by exposure to the intervention. A second purpose of assessment in instructional contexts related to the improvement of the program itself. We wish to assess instructional outcomes, again, reading comprehension in the example just given, for the purpose of revising instructional processes in the desired direction. These two purposes of assessment interact, often sharing the same sets of data collection processes and measures.

With both these outcome assessment purposes, the principal focus has been on the achievement produced by the intervention, what and how well students learn, or to flip the perspective, how well the intervention taught as a measure of its effectiveness. Recently, the focus of outcome assessment has been broadened in a number of ways: 1) to assess both cognitive and affective outcomes other than those intended by the intervention; 2) to include measures of attitudinal development and satisfaction; 3) to assess how students go about learning, their processes rather than their products. An additional but largely unsatisfied quest is to determine for which students, based on student individual differences such as cognitive preference, experience, and ability, various instructional interventions are most effective (Cronbach and Snow, 1977; Clark 1983).

But a critical focus is on the assessment of learning outcomes. The means to accomplish such assessment has been criterion-referenced measurement (CRM), and that is the major focus of this chapter.

Criterion-Referenced Measurement - Some Background

Criterion-referenced measurement has had many definitions. The merits of each and implications of different wording will later be discussed at some length. At the outset, we offer the reader a small sample of definitions which capture the range in the field.

A criterion-referenced test is one that is deliberately constructed so as to yield measurements that are directly interpretable in terms of specified performance standards (Glaser & Nitko, 1971, p. 653).

A criterion-referenced test is used to ascertain an individual's status (referred to as a domain score) with respect to a well-defined behavior domain (Popham, 1975, p. 130).

A pure criterion-referenced test is one consisting of a sample of production tasks drawn from a well-defined population of performances, a sample that may be used to estimate the proportion of performances in that population at which the student can succeed (Harris & Stewart, 1971, p. 2).

The history of norm-referenced achievement testing has been described in part by a range of scholars, each operating from a differing frame of reference (Nifenecker, 1918; Spearman, 1937, Cronbach and Suppes, 1969; Buros, 1977; Levine, 1976). The particular path of development of criterion-referenced testing is less well documented, although partial attempts at description have been produced by Millman (1974), Brennan (1974), Popham (1978), Hambelton (1978), and Baker (1980). Under contention, for example, is when criterion-referenced measurement (CRM) began. It seems to have two major sources: curriculum development inquiry and instructional psychology. Its early roots can undoubtedly be traced to Rice's assessments (1893), continued with Thorndike's experiments (1918), and Washburne's applications to school objectives (1922). The impact of Ralph Tyler's contribution cannot be underestimated, with his widely disseminated writing on curriculum development and evaluation (Smith and Tyler, 1942; Tyler, 1943; 1950; 1951). There is similar evidence, from the work of instructional psychologists, of the early development of CRM techniques for the assessment of instruction, for instance, films produced for World War II training (Hovland, Lumsdaine & Sheffield, 1949). In these early examples, content was sampled from the instructional universe of films, as is recommended currently by CRM specialists. The psychological bases of CRM was later exhibited in the experimental analysis of human and animal behavior (Skinner, 1958).

When reviewing the psychological roots of CRM, the source of nomenclature associated with CRM can be identified. For example, criterion itself simply meant a terminal or ending frame in a sequence of programmed instruction, where the response opportunity for the learner was unprompted (or without cues supporting the correct answer). Only later were such criterion trials aggregated into a criterion test of the sort Glaser described. Programmed instruction absorbed the attention of many psychologists concerned with changing student performance, who provided us with concepts such as task analysis (Gagne, 1965; 1977), performance level (Mager, 1962), and individualized instruction (Holland and Skinner, 1961; Lindvall and Cox, 1969).

CRM was first conceived to be a dependent measure for instructional sequences, sequences which were concrete and carefully designed. Thus the purpose of CRM was twofold: 1) to provide an operational definition for the skills developed by a given sequence, 2) to be used as a mechanism for formative evaluation (Scriven, 1967) as a way to improve instruction. The use of test information to revise instruction was a tenet of programmed instruction, and was also called developmental testing (Markle, 1967) or field trials (Lumsdaine and May, 1965). Of great importance, however, was that the test and instructional sequence were intimately connected, which made elaborate description of what the test measured unnecessary.

Early Applications

Fed by both the programmed instruction movement and the broader curriculum development and evaluation concerns of Tyler (1950) and Bloom (1956) was the movement in American education relating to behavioral objectives. Advocates of such objectives (Mager, 1962; Popham and Baker, 1968) argued that specification of goals allowed teachers greater efficiency in their instructional tasks as well as concrete means for assessing the success of their instruction. Although the movement often resulted in enthusiastic overspecification, with hundreds of tasks identified for a single course, the progressive refinement of the idea resulted in fewer objectives (to aggregate discrete objectives into clusters that were more sensible for learning and instruction). The emergence of more generalizable classes of behavioral goals and the recognition that the evaluation of these goals (testing) needed to derive from the clear statements led to the development of specification-oriented testing, or CRM.

From the Tyler tradition, and elaborated by the work of Carroll (1963), Bloom (1968), and Keller (1968), teacher-oriented notions of mastery learning developed. These models shared an important philosophic view, adopted, it appears, from the work of the programmed instruction designers: that student success was the shared responsibility of the teacher and the learner. Teacher training models were concomitantly developed, based on this point of view (Michigan State University, 1968; Popham and Baker, 1970; 1973). In addition, the curriculum development renewal, spurred by Federal support of regional educational laboratories and research and development centers (Title IV, ESEA, 1965), integrated Tylerian and programmed instruction traditions (see for example, products developed by the Southwest Regional Laboratory in California, or the Learning Research and Development Center, University of Pittsburgh). These instructional systems, whether purely programmed instruction, teacher-mediated, or comprehensive systems, depended for their evaluation on quality criterion measures. Thus, the initial utility of CRM was almost always as a part of an instructional system. The tasks assessed by CRM were circumscribed by the goals of the instructional system.

The Beginnings of CRM as a Field of Study

As shown earlier, critical definitions of CRM include the notion that performance is assessed relative to a particular task domain and that representative samples of tasks from this domain are organized to make a test (Glaser and Nitko, 1971). Glaser's work spurred the analysis of CRM as a measurement model rather than only as a part of an instructional system.

Early discussions of CRM, after Glaser christened the fledgling approach, struggled to contrast CRM from traditional testing theory. In their well known and referent article, Popham and Husek (1969) contrasted CRM and norm-referenced tests (NRT) on the basis of test development procedures, test improvement procedures, analysis and interpretation routines. NRTs were so named because their reporting procedures required that individual scores be transferred to a common scale and characterized as ranks in a distribution of scores. Thus, a score had meaning only in comparison to other scores in a particular distribution. Data were reported in terms of percentile, stanine, or quartile. It became gradually clearer to researchers that the norming process not only depended upon the selection of appropriate comparison groups of students, but also that it significantly influenced the development procedures of the test items themselves. The development procedure was bound by the requirement of performance variance to permit normal curve interpretation. Thus, early distinctions between norm- and criterion-referenced tests were drawn in terms of what was expected to happen to this variance after instruction. Because norm-referenced tests were developed to provide discriminations among individuals and relatively stable estimates of individual performance, instruction was expected to affect students about equally. The shape of a norm-referenced score distribution would not change as a function of instruction. Everyone was simply expected to move up a few notches (as the phrase grade-equivalent suggests). The relative rank of a student's score in a distribution was not expected to change. In contrast, criterion-referenced score distributions should alter dramatically after the treatment of related instruction. Before teaching, the pretest distribution might be homogeneously clustered and low on the scale for peculiarly obscure tasks, or for more general areas, randomly distributed; following instruction, it was conceivable for the great proportion of students to be achieving very high levels of performance, with relatively small variance. Before too long, researchers recognized the effect of reduced score variability on the utility of extant statistical procedures for examining test adequacy.

The Problem of Identity

Just as a young child probes the limits of his own identity and seeks to separate and distinguish himself from his parent, so did the writers in the area of CRM continue to seek to differentiate CRM from norm-referenced testing. Streams of articles attempted to describe what CRM was, including Popham and Husek (1969), Simon (1969), Lindquist (1969), Ivens (1970), Block (1971), Ebel (1971), Harris and Stewart (1971), Glaser and Nitko (1971), Emrick (1971), Cronbach (1971), Kriewall (1972), and Livingston

(1972). Much of these discussions focused on the model underlying CRM. There were two basic points of contention. First, the question was raised whether the term criterion meant a criterion set of behaviors, or essentially a task domain, whether it meant rather a standard or performance level, such as 70% of the items correct, or whether it was to be used as an external criterion, such as in criterion validity (Brennan, 1974). A second point of contention was how well specified were the domains from which the items were drawn. Some suggested that a CRM needed careful specification of both content and behavioral domains. The recognition of different degrees of specification led to analyses which not only contrasted norm and criterion-referenced tests, but also attempted to distinguish subsets of CRM, such as objectives-based, domain-referenced, and ordered sets. (See, for example, Denham, 1975; Sanders and Murray, 1976; Skager, 1975; Harris, Alkin, and Popham, 1973; Glaser and Nitko, 1971; Millman, 1974; Popham, 1978; Dzuiban and Vickery, 1973; Hambleton, Swaminathan, Algina, and Coulson, 1978; Berk, 1980; and Baker and Herman, 1983). The recency of some of the entries suggests clarity is not rampant in the field and, in fact, which concepts are subsumed by which appears to be a matter of personal preference by various writers.

Conflict

A good many of these articles and books attempted to distinguish between CRM and NRM by casting doubts on the goodness of one or the other (see, for example, Perrone, 1975; Haney, 1979; Ebel, 1972). Such doubts were easy to support on either side, for assessments of the quality of available commercial achievement tests, both norm referenced, (Hoepfner, 1971-1976; Haney, 1978) and criterion referenced (CSE Test Design Project, 1979) were generally negative.

From the literature alone, it is difficult to gauge the intellectual environment in which these discussions occurred, but in fact, a good deal of rancor was generated by contending advocates for norm and criterion-referenced testing. Within active memory were rather vitriolic exchanges between purveyors of the "upstart" form of assessment, the CRM devotees, and those firmly grounded in traditional psychometric theory. Debates were held at resesarch associations. National professional groups published resolutions in favor of one or another sort of testing, and then sometimes switched sides. A joint committee of the American Psychological Association, the American Education Research Association, and the National Council for Measurement in Education (1974) made an attempt to mediate differences (American Psychological Association, 1974). CRM advocates saw themselves as student and teacher oriented, interested in testing in the name of formative evaluation and the improvement of education. Norm-referenced test authorities held fast to the long and scholarly psychometric traditions upon which NRT was based. They could point to well developed concepts of individual differences, robust parametric analyses to assess the quality of their measures, and a thriving industry of users.

The sum of the criticisms of CRM by this group was that it was largely atheoretical nonsense. Should one review some of the early examples of CRM, such criticism is clearly appropriate. As will be detailed later, test construction in the name of CRM proceeded at a superficial level. Items were generated and reviewed under less than rigorous conditions (justified, of course, because the empirical analyses available to improve norm-referenced tests could not be directly applied and interpreted for CRM).

Social Context and the NRT-CRM Debate

One of the great ironies of this period of CRM development, the late sixties and early seventies, occurred as a function of the social reaction in American education. Precisely at the time CRM was emerging and differentiating itself under the banner of more educationally and instructionally relevant assessment, a strong reaction to technology of any sort took place. Both NRT and CRM advocates were tarred by the same brush by representatives of the counterculture, activists who rebelled against institutionalized testing and its attendant philosophy of logical positivism. Thus, CRM and NRT were thrown together as "the enemy", and distinctions between models of assessment were overshadowed by the general rejection of "irrelevant" and competitive educational activity. These reactions, scholars avow, were in part caused by social disruption, the limited success of the Great Society (Aaron, 1980), and evidence of the perversion of public political power.

At the same time, and causing additional conflict in the practical world of education, was the increasing public attention and support of testing (Atkin, 1980). The evaluation requirements attached to Federal categorical aid programs spread the amount of testing throughout the nation. The interpretation by the courts of test data, such as reported in the Coleman study (1966), the trends toward statewide achievement programs, and the development of school leaving examinations as a criterion for high school graduation (Pipho, 1978) raised the testing stakes. What had started as an academic squabble between educational psychologists grew to an issue of considerable proportion in public policy. As the testing issue became more visible, and involved life choices of individuals, so did the need to identify problems in the testing field become more urgent. Consumer advocate groups (such as Nader's) attacked testing institutions, questions regarding test security were raised concomitantly (Haney, 1978), teacher organizations presented forceful points of view (NEA, 1979; Ward, 1980), contention was fed by court cases and legal analyses of tests were issued (McClung, 1978). Another broad irony is that most of these analyses of test properties were based on work of psychometricians, a professional group with relatively little school experience and almost no involvement with instructional programs.

Especially noteworthy in reviewing the development of CRM is that only rarely were the core philosophic distinctions between NRT and CRM clearly articulated. Bloom (1968), in his classic article on mastery learning, pointed out the difference in expectation such a model could make for children and outlined some of the benefits of allowing learning time rather than student competency level to vary. One clear consequence was the sharing of instructional responsibility by teacher and student. Not yet solved, however, are the practical difficulties of implementing such an idea in the face of continued social and financial pressures in schools. These difficulties include problems associated with realocation of resources to students who require more time, the nature of shared responsibility in the face of high student absentee rates, and the tendency for mastery to be set at lower rather than higher levels (Baker, 1978).

Test Design for Criterion Referenced Measurement

When one imagines what ought to be in a section called test design, a prominent contender is how to make a test, that is, the nuts and bolts of actual item writing and test assembly. While such activity has rarely been regarded as at the higher end of the intellectual continuum, nonetheless rules, procedures, and routines for test construction have been developed, for use by either the professional test builder or by teachers. In this section, some contrasts will be presented between test construction activities and test design efforts, the former characteristics of typical achievement test development and the latter examples of test development in CRM.

Norm-Referenced Test Development: In Brief

Certain steps in achievement test construction were developed in traditional practice. It should be emphasized that the routines were created 1) to assure a broad representation of item and content types; 2) to avoid gross technical error. The major burden of test development for norm-referenced achievement tests (NRT) fell on empirical analyses.

Typically, in NRT, a general content-behavior matrix was first developed, so that test items could be generated to tap the full range of topics and eligible response modes. Then items were reviewed to assure that they did not inadvertantly cue the learner to the correct answer, that the length and syntax of response options were comparable, and that the correct answer was keyed accurately. These items were also inspected for content quality and screened for obvious technical errors. Most important in test development processes, however, was the use of empirical procedures to determine test quality. Techniques such as item analysis, reliability estimates, and quantitative indicators of validity were created to help the test item selection process. These techniques were based upon parametric statistics used by researchers in analyzing experimental data. Such techniques depended, as did certain experimental research models, on classical notions of science: predictability and control.

Underlying empirical test refinement practices was a relatively simple idea. A norm-referenced achievement test was to measure a general ability, pertinent to an area of knowledge or skill. The underlying "explanatory concepts...accounting for test performance" were called constructs (Cronbach, 1971). An individual's performance included chance exposure to relevant experience, broadly aggregated, as well as to in-school or other purposive instructional experience. Constructs, definitionally, required more than one measure. Performance on any single test measuring a general construct (such as reading ability) was thought to provide a relatively stable estimate of an individual's performance when compared to other similar individuals. The role of change (as in learning due to instructional exposure) was noticeably unclear. As such achievement measures were to assess important dimensions formulated as constructs, the argument ran, then they should not be reactive to relatively small variations in the learner's total experience, for instance, whether or not a child received a particular one month reading comprehension program. Such a model was almost universally accepted and maintains strong and eloquent supporters (see, for example, Ebel and Anastasi, in Schrader, 1980). They describe a view of achievement as a developed ability, with the other end of a continuun anchored by aptitude (the capacity or predisposition, without the relevant experiences). This notion of achievement was supported by statistical analysts who conceived of testing in terms of prediction. Changes in test score from occasion to occasion were formulated as unreliability or error (see, for example, Harris, 1962) by such methodologists.

Certainly, no one worries much about models underlying test construction or any other human endeavor when certain conditions hold: (1) performance looks good; (2) significant decisions do not hinge on the model's products; and (3) a body of prestigious support is available for the practice. Such was the comfortable status of norm-referenced achievement testing for many years. Measures now show a less than rosy view of student achievement, and explanations for declines have not been satisfactory (Wirtz, 1977). Decisions about admission to professional schools, coveted undergraduate institutions, and even the award of the high school diploma increasingly depend upon test performance. Obviously important, perhaps, is the lack of scholarly consensus on the quality and utility of achievement measures. Because these issues focus attention on the effectiveness of schools, a different philosophy about education has developed vocal, if not always coherent, support. That view is also simple: that schools exist to produce change, in other words, specific learning. In this view, change is not regarded as score unreliability, but is itself the most desired product of education. One should note the level on which discussion of this issue has occurred. Secretary Joseph Califano, then head of the Department of Health, Education and Welfare, made a public statement where he avowed that the federal government wished to reduce the predictability of performance based on socioeconomic or race classifications (1978). Since relationships in status on these demographic variables and standardized test performance run very high (between .60 and .80) depending upon the reliability of the test, and student performance on similar tests correlate, over time, at .80 or higher (Bloom, 1980), one may infer that this statement challenges the test development community to build measures able to detect effects of educational practices within the school's control. In contrast to earlier formulations, change is to be valued over predictability. This perspective shift has great implications for test construction. Procedures used to develop measures of traits thought to be essentially stable over time are not the same ones that should be used to create change-responsive outcome measures (O'Neil and Richardson, 1977).

Specifications of Tasks

CRM developed, it was earlier noted, out of two traditions, each actively promoting change: instructional psychology and curriculum development. Both of these sources, although from different governing frameworks, hit upon the practice of specifying objectives or goals for change. The practices in CRM development grow from the answers to various questions related to this specification or description: What is specified? At what level of detail? Where do the specifications come from?

In the earliest days, specification of tasks for assessment were thought to flow very nicely from a clear statement of an instructional objective (Mager, 1962). Although these objectives could be developed to cover course-level material, they were usually created for shorter units of instruction. The belief was evident that, once figuring out how to state an objective clearly, test development would be a cinch. In rules designed to help in the assessment of educational programs, Popham (in Baker and Schutz, 1968) suggested that the critical measurement issue was the classification of forms of stimuli and responses. As an early advocate of diverse forms of measurement, Popham classified assessment tasks into four cells: (a) student behavior could be either process (throwing a ball) or product (test paper); (b) elicitation conditions could be either formal (school) or natural (out-of-school or surreptitious). Additional writing around this time focused on how specific the specification needed to be for the assessment ("to take a test" was a negative example, considered much too vague). Also of interest were conditions under which the test was to be taken (time limits, extra materials) and ways of establishing desired performance standards (such as 75% correct). While Tyler and others since had noted that an objective consisted of both behavior and content, a good deal of early attention in objectives-referenced measurement was devoted to specifying behavioral requirements and very little in developing the content parameters. Good items were thought to match the behavioral statement in the objective.

The Problem of Content

In the absence of routines for specifying the what (content) of testing in favor of the how (test behavior), two rather different modes of practice developed. Test items were selected or rejected on the match between the objective statement and nuances of the test taker's behavior (was the student directed to cross out a letter when the objective called for a machine scored blacked in response?). In one mode, content was left to vary freely without any specification ("important mathematics concepts" or "American novels"). In the other, each particular content unit was specified ("In the play Othello, identify..."). The trade-offs appeared clear: in the first case, the task was cast in a generalizable form, for almost any particular content would be eligible for inclusion in the test. In the second, particularization of content allowed for highly targeted instruction and congruent testing, but forsook generalizability. Discussions of the merits of these trade-offs, generalizability vs. specific content, were held in workshops and training sessions of the American Educational Research Association during years from 1967 to 1973. However, real confrontation with the content of tests, that is, the subject matter areas to be assessed, was generally limited. Although there were analyses of new curricula, new math, the process-oriented new sciences, the new linguistics, such were not specifically analyzed for their utility in developing performance-oriented instruments. Content people were generally too "soft" for the hard edged requirements of behaviorism, and remarkably few content specialists were interested in testing specifically. During the mid-sixties, an impetus for a new view of content in objectives-based testing was needed.

Domain-Referenced Achievement Testing

The work of Osburn (1968) and Hively (et al., 1968) provided that impetus. Using a model developed from set theory, Hively described the identification of a universe of content and behavior, a domain. Hively demonstrated that broad classes of performance could be assessed by using algorithmic rules to generate items. This domain could then be theoretically sampled to yield representative instances of test items. Performance on the sample would allow the estimation of performance for the larger content/behavior domain. Hively, in refinements with colleagues (1973, 1974) demonstrated how a technology for domain-referenced test (DRT) generation could be developed. He suggested the use of an item form, or shell, that included basic behavioral requirements. Into this shell could be inserted replacement content instances, substituted from the "universe". A simple example of an item form is the addition problem:

$$x + y = \underline{\qquad}$$

where x is any two digit number and y is any one digit number. While the item shell might be changed to:

$$\begin{array}{r} x \\ \underline{+y} \\ ? \end{array}$$

the content parameters would be identical. Two digit and single digit numbers were to be added. Any members of that set in the specified combination might actually show up as a test item.

Hively's suggestions had great impact for a number of reasons. First, as described earlier, there was dissatisfaction with extant test development processes in the field. While there was recognition that available empirical procedures were inappropriate to apply to new outcome measures, no alternative procedure had been agreed upon to produce quality test items. Hively's work probably also indirectly capitalized on the widespread knowledge of Bloom, Krathwohl, and colleagues' (1956; 1964) efforts at taxonomic organizations of educational objectives. The term domain used in these works was understandable to all. An additional explanation for the success of Hively's ideas was his development and demonstration of domain-referenced achievement testing in concrete form. He provided a real example to researchers in the field, an example couched in a theoretical context but which had practical implications. He had actually created test items using such procedures.

Forms of Items Forms

Hively's rules for the creation of items included the specification of the formal of the item, the rules for generating the stem, the response alternatives, and the directions. When fully explicated, his item form directions appeared detailed and formidable. Such detail was clearly required in order to develop unambiguous item domains. Yet his procedures, because of their sophistication, seemed designed principally for use by a team of item writers. Baker's adaptation, reported in Hively's book (1974), focused on specifications as they might be modified for teachers and others familiar with behavioral objectives. The elements of a domain specification included a statement of the objective, the content limits, the wrong-answer population (for multiple choice tests) or response criteria (for production tasks), the item format, the directions, and a sample item. Popham (1975) further modified domain specifications to what he termed an amplified objective. In his scheme, stimulus attributes and response attributes were to be specified; however, distinctions between the behavioral and content requirements of the item were not made. The Popham and the Baker adaptations represent less rigor than the Hively approach, but were justified in terms of likely comprehensibility to teachers and instructional designers. At the outset, these approaches were applied to single domains and the problems of creating tests across a number of related domains was not addressed.

Hively's work was particularly important because of its connection with instruction. Unlike the curriculum development people, who saw specification of objectives and measures as one of the first steps in the process, Hively had directly referenced his efforts to extant instruction. He used content generated by lesson writers as the primary source for the creation of his item domains. Similar to the way in which programmed instruction linked its criterion-frames to instruction, so Hively's item forms were linked to the concepts in actual lessons. Although his work was extended by Popham, Baker, and others to the objectives-instruction-assessment sequence, his ideas remained firmly grounded in instruction. Domain-referenced testing (DRT) immediately formed a new category of criterion referenced measurement, and writers described applications in teacher training, program evaluation, and accountability (see, for example, Hively, 1974; Harris, Alkin, and Popham, 1974).

DRT generated fodder for intellectual rumination lasting well into the most recent period. Questions were raised, and almost endlessly discussed, by Popham (1978), Millman (1974), Hambleton (1978), Baker (1978), Brennan (1974), Harris (1980), Haladyna and Roid (1978), Nitko (1974), and Anderson (1972). Numerous problems in DRT were identified and lists of unresolved problems published in 1974 appear to continue in that status.

Problems of Domain-Referenced Testing

Among some of the early problems associated with DRT was the attempt to deal with content parameters outside the field of mathematics and science. Although it was very clear how one might go about generating a set of parameters or generation rules for computational questions, doing so in the liberal arts appeared to be a messy process. Hively's procedure was based upon an algorithmic approach to content selection. Thus it was especially applicable to content areas that had well-defined structural relationships, such as an early example of DRT in a linguistically oriented reading program (Baker, 1968). In this example, a specific set of rules governing content, such as syntactic and spelling rules, allowed for the explication of a universe of content and the compilation of tests that sampled the defined universe.

The attempt to apply DRT to other subject-matter areas were many, and included social studies, writing, English literature, the health sciences, and reading comprehension. A major fact soon became evident: few subject matter areas had sufficiently well-defined structures to permit the use of algorithmic approaches to content generation (Landa, 1974). In the absence of sufficient clarity in subject matter fields, would-be users of DRT fell back on an alternative process. Their choice was to define the parameters of content operationally themselves, without reference to any subject matter analyses. They would decide, for example, that four causes of economic decline existed, list and define such causes, and develop examples of each. A DRT could then be created by selecting an appropriate range of examples. This method was clearly vulnerable to charges of both arbitrariness and curriculum control. Defenders of this strategy pointed to the void in current practice and suggested that this technique was preferable. As a coincidence, Gagne (1977), in an audiotape developed for AERA, discussed two forms of concept learning. The first type, _concrete_, were those derived from perception. The second category of concepts were those he called _defined_ concepts, where the instructional designer (or test writer) would explicate the dimensions of a concept and the learner would discriminate examples or generate instances based on these defined or agreed upon limits. The use of such defined concepts supported the DRT content specifications. A large and unresolved issue remained: who was to decide on the arbitrary features of a defined concept. No satisfactory and practical answers have been suggested, from the measurement community beyond the usual discussion of constituencies and judgment by reasonable persons. The advances in cognitive science, however, presage improvement in specifications. Both cognitive skills and precise content representation may contribute to resolving this issue (Curtis & Glaser, 1983; Baker, 1985).

A second major problem was what to do in cases in which the subject matter itself defied algorithmic definition, even an arbitrary one, in a case such as literature. While it is conceivably possible to specify arbitrary rules for generating examples of lyric poetry, the exercise seems relatively futile because of the variation of examples within that literary genre. Taking a cue from Hively, some DRT writers identified domains not by _generation_ rules (for all possible instances) but by _enumeration_ of a limited set (for instance, poems 1-9 found in Smith's anthology). Such a tactic reduced the power of DRT to claim estimation of a total domain (such as poetry), reduced the likelihood of generalization (that perhaps performance levels would be similar from poem to poem), but preserved the "fairness" with which items might be sampled by circumscribing the set of content to that contained in the particular anthology. Thus, at least, students and teachers and test writer would know what content was fair game for testing.

Another fall-back tactic for content specification was to define by illustration and axiom a set of content. Hively provided the example of the frontpage of *The New York Times* as a content set for assessing reading comprehension. Clearly the explication of generation rules or algorithms for content such as *The Times* is beyond both the funds and attention spans of researchers. In another example, the operational definition of a *clear sentence*, including forms of reference, semantics, and soon, similarly over-complicates a domain more intellectually accessible by example. As provided in any number of style handbooks, clear sentences can be clearly contrasted with unclear writing. The *rules* are more efficiently perceived in the examples themselves, rather than exhaustively written. Again, this form of specification, while short of the purity of item generation rules, clearly communicates to teacher and learner what is to be tested and what should be learned.

The problem of the completeness of content domain specification can be recast as a problem in automation. How fully automated should DRT's be? The extent to which test item writing can be fully automated is presently unknown but approximations using domain specifications or syntactic rules have been attempted. Bormuth (1970) provided essentially linguistic transformations to permit the generation of test items. In a series of studies to assess the automaticity of item writing, Roid and Haladyna (1978) were surprised that item writing "subjectivity" was not removed by the provision of rules to two item writers. In another study using prose passages, Roid, Haladyna, and Shaughnessy (1979) found some algorithmic practices controlled item writing production. The study supported the importance of linguistic analyses of items in addition to other specification matching routines. This study was also limited, however, by the use of only a few (four) item writers. Undaunted, they continued (Roid, Haladyna, and Shaughnessy, 1980) with six item writers directed to use linguistic vs. subjective (match with an objective) strategies. Although lengthy analyses are provided, the item by item writer interaction suggests that item writer behaviors were not sufficiently effected. The authors posit the need for further trials with more empirical tryouts. However, tryouts under conditions of good, medium, or rotten instruction would likely affect the resulting data set. Baker and Aschbacher (1977) achieved considerable success in controlling item production through the use of rules. The automation problem has not been discussed in most research in this area. The use of the computer to automate item writing routines has been less well-developed to date than one might hope, with only relatively simple content substitutions used. Millman and Outlaw (1977) conducted a project in this area and Finn (1978) reported on multiple-choice item generation. Hsu and Carlson (1973) earlier used the PDP-10 system, and other automated experiments involved efforts by Olympia

(1975) and Fremer and Anastasio (1969). This work needs to be linked and made more relevant to the content parameters of domains. Perhaps availability of better natural language processing options would improve computer utilization in this important area (see Frase, 1980; Freedle, 1985).

New Approaches to Content Specification

While computer technology has long been employed to score and to administer tests (Dunn, Lushene & O'Neil, 1972; Hedl, O'Neil & Hansen, 1973), its exploration may have some utility in the content specification problem of domain reference achievement testing. Specifically, the development of expert systems provide an opportunity for specific knowledge domains to be identified, structured and incorporated into computer software. Basically, these approaches focus on the problem of representing expert knowledge and its relationships in algorithms that the computer can use (Buchanan, 1981). Modelling knowledge via expert systems have, by and large, focused on relatively narrow knowledge domains, such as subtraction (Brown & Burton, 1978), but efforts have been made to attack more complex areas, such as computer programming (John & Soloway, 1985), infectious diseases (Clancy, 1982), story generation (Dehn, 1981) and understanding narrative (Dyer, 1982, and Fredericksen & Warren, 1985). Research is also underway to develop procedures for less well defined areas, so called fuzzy content (Spiro, 1984) where content does not fall into mutually exclusive categories. The techniques used to represent knowledge developed for AI expert systems could be used in the vexing problem of assuring full content representation on tests.

Quality Control

Another nagging question about DRT is how one knows an item is a good instance of the set. Most writers suggest some judgment scheme, usually matching the item realistically against characteristics explicated in the specifications. Research on this problem has demonstrated that raters may make their discriminations on superficial item features; for example, does the number of response alternatives in the item match the specifications? rather than on the more difficult issues of cognitive complexity or content appropriateness. Some research has been conducted relating to the need to provide guidelines for such judgments (Polin and Baker, 1979).

Using defined concepts and operating from an instructional perspective, rules and routines for matching instances with classes have been developed by Markle and Tiemann (1974), Tiemann, Krocker, and Markle (1977) and Tiemann and Markle (1978a,b). Merrill and Tennyson (1977) have also provided excellent analyses and examples of processes needed to match examples of concepts to specifications or concept definitions. Because this work takes place in the context of instructional rather than test design, these authors have received less than their due recognition for contribution in the testing field.

Of the research conducted on providing guidelines for judgment in a test design context, Hambleton (1980), Haladyna and Roid (1977), Baker and Quellmalz (1977), Doctorow (1978), and Polin and Baker (1979) have made contributions. Set theory, or more particularly the concept of fuzzy sets, has been applied in this research to estimate the degree of congruity between an item and its specification. This research demonstrates the futility of using obvious and superficial indicators (such as the number foils in the specifications); and factors such as level of cognitive complexity and related linguistic features were highlighted as needing more study. A number of writers have reported training efforts undertaken to teach specification - item matching (Merrill, 1979; Tiemann and Markle, 1978; Hambleton and Simon, 1980). Baker, Polin, and Burry (1980) have developed training materials designed to teach the rudiments of DRT judgment to teachers and to graduate students. Such training seems to be required before individuals can match test items with their specifications. Secolsky (1980) makes the argument that students must be able to match relevant items with their generation specifications (i.e., to label concepts, to demonstrate that the items cohere). This rather demanding requirement might be acceptable if students were first trained specially in identifying the critical attributes in DR items. In the absence of such training on relevant dimensions, students might group items under true, covarying but instructionally irrelevant features (such as sentences starting with the letter T). In the development of the review process described earlier investigated by Polin and Baker (1979), the critical issue was training item classifiers on instructionally relevant item features.

The foregoing problems that deal with the match by inspection of specifications and items represent what Bormuth (1970) calls problems of item-writing theory. His second category deals with item-response theory, or more accurately empirical indices used to substantiate the existence of a domain. Millman (1974) also attempted to distinguish between problems of item selection which were judgmental and those for which empirical data were necessary. Popham (1978) also distinguished between descriptive validity (that is, does the item fit its specifications and are those specifications clear?) and functional validity (does performance classify the student as anticipated?). Early interpretations of the DRT process included high expectations of item homogeneity, as discussed by Nitko (1973). The idea was that item difficulties and variances for items produced by DRT procedures should be similar. Items were expected to cluster together (Baker, 1971; Macready & Merwin, 1973; Stenner & Webster, 1971). Cronbach (1972) discussed procedures where individual item writers would be able to produce items which resulted in similar empirical characteristics. Although this demand for homogeneity has diminished in the light of actual data sets, one may still be troubled by the idea that item performance, particularly one developed by DRT procedures, was assessed in the absence of clear documentation of the instructional conditions preceding its use. A similar issue may by looming for the advocates of new empirical procedures thought to obviate the requirement for meticulous matching of specifications with items. The Rasch model (Wright, 1967) has been put forth and scooped up by users of CRM as an empirical solution to the issue of item quality. What is still unclear, however, is the extent to which this model, and in fact other latent-trait (Boch, Mislevy, & Woodson, 1982; Boch, Gibbons, & Murchi, 1985) models are robust in the face of highly targeted instructional interventions. Research by Roid and Haladyna (1980), albeit exploratory, does not lead one to expect good news. Somehow empirical analyses, combined with judgment of specification to item matches, conducted under known instructional interventions, will be necessary before we can uncritically adopt solutions such as the Rash model proposes.

Matching items to specifications or the generation of item sets according to specifications is based on a pigeon-hole view of the relationship of given items to a domain. Each item would be sorted as it fits according to the exhibition or absence of N features explicated in the domain specification (Choppin, 1980). It is altogether possible that limitations of item writers, subject-matter structure, and technology will conspire to promote alternative, perhaps supplementary models to DRT. One such area of analysis involves the linguistic features of test items, beyond the readability indices presently computed. A similar technique area once again ripe for exploration is the area of facet analysis and concept mapping (see Engle & Martaza, 1976; Gutman, 1969; Harris, 1976; Beck, 1978). The improved natural language processing capacity of computers may also enrich our DRT technology. One principal incentive for such work may be the need for procedures for the development of access and retrieval routines for computerized item banks. Such techniques could easily influence item development and review processes and result in significant improvement.

The foregoing discussion pertains principally to the technology of comparing sets of generated items with their parent specifications. Only oblique discussion has hinted that the content and behavioral requirements themselves might requirement review. Along what dimensions might specifications be judged? In much the same mode that goals and objectives were to be judged by relevant constituencies, so too might domain specifications be reviewed for relevance and importance in school learning. Some critical questions still need research before we could even begin to open the review process to less technical participants.

For example, how big is a domain? The answer was at first thought to depend upon empirical data (to wit, a domain has items that cohere), but as strict expectations for item homogeneity faded, so have guidelines for the restrictiveness of domains. How much complexity in a domain? Are homogeneous response modes required? Does a domain include the task to be tested as well as relevant sub-tasks in an identified skill hierarchy? Do such subtasks need enumeration or do they also require verification empirically? How are domains organized with respect to one another? In parallel? By content area? In more than one way? How are task requirements best determined? As pointed out, for the most part specifications have grown from the analysis of content areas and rather gross behavioral requirements. In some cases, instruction itself has generated the parameters. What should be the relationship of instructional analyses to domain design?

Integration of Testing and Instruction

The relationship of domain specification to instruction is an area which might profitably be addressed. Certain models start with instruction or content (see Hively, et al., 1973) and reference the domain to that set. Others start with the test specifications, and then develop instructionally relevant learning opportunities (see Rankin, 1979). Thus from given domains, test specifications, item pools, and instructional practice exercises are generated. This system does not completely specify all instruction but it is designed to integrate some aspects of domain design with testing and instructional functions. In mastery learning (Bloom 1969; Block, 1971), a natural oscillation between instruction and testing occurs.

Researchers are presently at work attempting to find ways to connect instruction and testing at deeper levels than in the past. Rather than developing tests to reference extant instruction (see the Proficiency Verification System, SWRL) or to map extant tests on instructional texts (Floden, et al., 1980; Porter, 1980; Montague, Ellis, and Wulfeck, 1983), ways to unify the design of test and instruction should be explored. Initial development of this sort has taken place with the creation of Project TORQUE (Schwartz and Garet, 1982), a math program where exercises serve almost indistinguishable functions of teaching and testing. The cognitive specifications for such a set of activities probably needs additional refinement. Frase (1980) has worked on the integration of testing and instructional domains using computerized language projects, and the research in writing assessment (Baker, 1982; Baker, Quellmalz and Enright, 1982; Purves, et al., 1980; Quellmalz, 1980) has potential for a similar sort of unification. Such a merger of instruction and testing will not come about easily. For one thing, it violates our traditional patterns of thought. Brennan (1974) expresses little patience with those who continually blur the distinctions between testing and instruction and impede, he believes, serious progress in either. On the other hand, a scholar as prestigious and traditionally grounded as Harris (1980) has seen the need to integrate testing and instruction complexes.

Most writers on instruction and testing have, in recent years, seen tests leading instruction, as in "teaching to the test". Mastery learning made a great contribution towards the integration of instruction and testing in two ways. First, the intervals between instruction and tests were reduced and made more frequent. Second, they were individually tailored for individuals (Rudner, 1978) or groups. Adaptive testing, using the computer to administer tailored items is a current example of this approach. Thus, the pattern was changed from formal and extended periods for testing and instruction (courses with only one mid-term examination and one final examination) to more flexible and naturally occurring events. But in the hearts and minds of many, instruction is still the treatment or intervention and testing is still the dependent measure.

For an analogous example, recall some of the early processes in the attempt to teach young children to read. An important and persistently difficult skill was the blending of initial consonants and phonograms, so that when a child was presented with the elements T and AN, he or she could pronounce TAN. For some reason, instruction focused on reducing the interval between the pronunciation of elements. By shaping the child's behavior so that the time between the pronunciation of T and AN was very short, the child would come, it was thought, to understand the process of blending. In fact, no such insight typically occurred. Children showed remarkable resiliency and ability to keep the two elements separate, even when the time between them was essentially eliminated.

Children did learn to blend easily, however, when the focus was not on reducing the time interval, but in changing the framework in which the blending instruction took place. In early experiments (Baker, 1968), children were taught to first understand the unified outcome that was desired, that the units had meaning, and blending was a process similar to saying SAND - BOX. When presented with T + AN no hesitations occurred and blending skill became well developed. Similarly, a new dimension must be found to underlie both testing and instruction so that these functions lose their uniqueness. Of great promise is the work in cognitive psychology, which, if united with theories of content structure and language, could allow the generation of experiences useful to develop and assess, in a piece, the desire outcomes of schooling. An excellent analysis of the future has been described by Curtis & Glaser (1983).

Narrow Definition of Testing

As we discussed, most individuals writing in the field assume a test is a paper-pencil vehicle, usually in multiple-choice format. They also seem to assume 1) that the test has one correct answer and that other alternatives are no more than "foils" to the right answer; 2) that the test is kept separate from instructional activities; and 3) that the present practice is probably most efficient.

There is only occasional mention of "performance testing", and a few writers grope to find words to distinguish other than multiple-choice testing. They use words like appraisal, evaluation, assessment, their Roget's litany, to avoid the constrained "test" connotation. In reflecting on this review, the reader would be wise, we believe, to make the effort to break out of a confined view of testing. The research should be judged as it could or might be expanded to generalize to formats of the sort listed in Table 1.

Table 16.1

Test Format Options

Format	Examples
1. oral language	Formal speeches, conversational facility
2. written composition	essay examinations, expository analyses, description, poems
3. physical activity	diving, tennis stroke
4. creative production	art, carpentry
5. technical exhibition	piano recital

Evaluating Instructional Technology

One of the most useful options in considering outcomes of computer-based instructional interventions is to use the technology of delivery as a means of collecting information related to student outcomes. Not only can the computer deliver tests that are embedded in instruction but it can also tabulate indicators of other instructional outcomes. For example, in the evaluation of a set of computer-based instruction, the latencies of student responses, the numbers of options they selected, the frequency with which they selected harder problems can be incorporated as an additional outcome measure of program effectiveness. In some sense, these indicators involve using processes as outcomes. The student is encouraged not only to improve his level of attainment but his fluency and exploratory behavior as well. Other automatically recorded information can provide indices of student attitudes - for instance, persistence and attention.

It is true that scholars working in the measurement area are moving toward a fuller concern with the understanding of student learning processes leading to particular levels of attainment. For example, Linn (1985) describes a measurement approach that tracks metacognitive processes learners employ as they encounter new reading requirements. Furthermore, Shavelson & Salomon (1985), undertake a study of the relationship of the symbol system in which the test is conveyed and the cognitive processes students use to develop their responses.

The availability of new computer technology for assisting in assessment problems has both positive and negative sides. On the one hand, it can encourage the intergration of assessment into the instructional context, so that it is more representative, less ceremonial, and less artificial than tests of the past. On the other hand, our analysis of what has been happening to testing as implemented in new technology is relatively negative. Short-answer and multiple-choice formats abound, and as a result, the performance tested is at the lowest common denominator possible. Tests, however, only mirror the approach taken toward instruction. When tests are molecular and discrete rather than integrated and comprehensible, one can make inferences about the quality of thought behind the instructional development effort even before seeing the data. We expect to see in future assessment, expansion and integration: where a common database can be explored to make inferences about performance, levels of attainment, relationships to individual differences, cognitive processes, and attitude development. Such an integrated database approach is possible now. However, as long as assessment continues to be regarded as the stepchild of instruction, a necessary evil for reporting requirements, rather than an integral instrument in the design of instruction and the teaching of students, few developers take the risk.

Integrating Assessment into the Evaluation of New Technology

While the foregoing sections have focused on assessment and the measurement ideas that underlie it, it is important to place concern for outcome measurement in context. What else needs to be included in the assessment of instructional technology that is especially relevant to the technological character of the innovation? In other words, what else needs to be addressed beyond measures useful for the assessment of non-technology based instruction? Let us turn, for the conclusion of this chapter to the issues related specifically to evaluating technology. Our assumption is that the best ideas posited for the measurement of instructional outcomes will be necessary but not sufficient for this evaluation task.

Assessment, and the evaluation processes which support it, is represented to be a productive mechanism for the improvement of educational systems and products. And there is hard evidence of the utility of evaluation in actually improving technology-based products and efforts in instructional development (Baker, 1972; Rosen, 1968). Assessment is known as well to contain a strong negative potential. Evaluation can identify weaknesses in such a way as to inhibit exploratory behavior and risk taking on the part of researchers and developers. Playing it safe may be seen to be the winning strategy. Evidence of evaluation utilization studies suggests that when the focus of the assessment is classification or accountability (good vs. bad; useful vs. wasteful), the openness of R&D project personnel to evaluation processes is inhibited. Formative evaluation, on the other hand, is evaluation whose specific function is to identify strengths and weaknesses for the purpose of improving the product or system under development (Baker, 1974; Baker & Alkin, 1973; S.M. Markle, 1967; Baker & Soloutos, 1974). The trick, of course, is in determining what should be studied, in what context the evaluation should take place, when evaluation processes are most useful, and in skilled hypothesis generation about what improvement options logically and feasibly may be implemented. In addition, the identification of weaknesses (no matter how benign the intentions of the evaluation may be) creates a documentary trail that might be misused by project managers or funding agency monitors.

These issues take on special dimensions when the evaluation addresses the effectiveness of new technology. All technology development of necessity focuses on the initial problem of system operation: can the envisioned delivery system work at all, as opposed to the refinement of what the system's merits may be or what effects might be planned or imagined. Outcome assessment is often a deferred goal. When dealing with emerging technology, the boundaries between technology development and science become especially blurred. The creation of technology may be a pleasant side-effect for the creator, whose perception of the main task may be knowledge production, rather than instructional effectiveness. Intellectual exploration is a premium for new technology development, and assessment processes can be seen to inhibit or be irrelevant to invention.

Recent writing in the field of evaluation planning has emphasized a stakeholder perspective in evaluation implementation. Simply put, this means that interested parties must have an opportunity to understand and to shape the nature of the evaluation questions and methods so that they will be more invested in the process and more apt to use any results generated (Bryk, 1983).

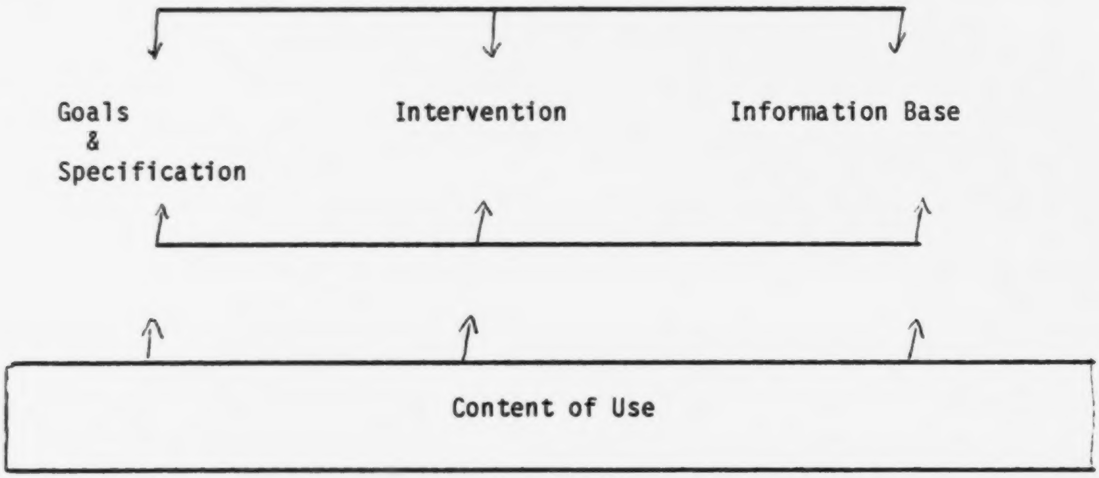
With this discussion as context, a special model of evaluation can be designed to be adapted especially to the problem of new technologies. Briefly, we will detail the features of this model, as applied to new technology, a particularly difficult area characterized by weak boundary conditions between research and application goals.

A Model to Assess Technology

The model underlying the formative evaluation of technology is composed of a minimum set of pieces. They include the goals or specifications, the intervention, the context of use, the information base, and the feedback loops. Figure __ displays this model linearly, but it in fact could be arrayed in a circle or three dimensionally. Points of entry to the model could vary depending upon the designer's committment to prespecification of outcomes, for instance. Or the extensiveness of alternatives could differ, with some designers interested in contrasting alternative instructional treatments and others interested in a broad array of outcomes, including attitudinal and social goals beyond those detailed in the system specifications.

Figure __

Model to Assess Technology



Desired Information Features and Functions for a Model to Assess New Technologies

Below we provide a list of four desired attributes for a model to assess new technology. These characteristics respond to particular attributes of technology development. In brief, these include weak boundary conditions between research and application goals of the developers; levels of risk in technology development; and the constant pressure to develop and sustain management support and necessary resources to complete the tasks of interest.

The information must provide an enhanced documentary base for the processes of new technology development. A characteristic of new technology is lack of documentation describing the process leading to the development of the system or product. The purpose of a strong documentary base is to provide the trace of developmental processes so that the field can improve overall. Aggregating across a series of case histories of projects can allow the inference about productive strategies to be made. In addition, a good documentary base can inform about dead-ends in substance as well as in developmental processes. Since most R&D reporting is based upon positive findings, it is difficult to avoid useless but unreported paths.

This lack of documentation exists for a variety of reasons. First, the process of early design of technology is complex, iterative, and non-linear. All of us are familiar with documents of development which retrospectively rationalize and make "neat" processes that are chaotic, or at best, hard to track. Furthermore, the metacognitive awareness required of designers to document their own processes while at the same time working on problems of interest presents an almost insurmountable attention burden, even if there were predisposition on the part of the research and development personnel to do so. Solving the problems at hand appears to be more important. Contributing to an abstraction such as R&D processes attracts less compelling energy, despite the intellectual apprehension that the field overall can be improved by a "lessons learned" perspective. Another inhibition is the precedence of proprietary knowledge, well known in the private sector, but of potentially increasing import in a public R&D environment characterized by competitive procurement policies.

In an attempt to meet this overall goal in instructional technology, some case histories were prepared 20 years ago (see D. Markle, 1967) and an historian was even on the payroll of another large R&D facility. But these persons can be as pestering and diverting as media reporters, trying to get the idea of what's going on without true understanding of the processes involved. In new technology development, the problem is obviously exacerbated.

Fully participating formative evaluators provide another model, however, if they are linked early on in the development process, and if the R & D management and staff understand the intent is to assist as well as to document process.

The information must use state-of-the-art evaluation methodology, including both quantitative and qualitative approaches to measurement. One of the reasons evaluation processes have been received with healthy skepticism is that they appear to be so content-free, on the one hand, and methodology-driven, on the other. The history of assessment and of evaluation, as in any new mode of inquiry, is replete with "new" models that propound a particular methodological view of the world. A good deal of the discredit done to evaluation has occurred with the support and consent of its most famous practitioners, who advocated one or another highly quantitative design and analysis method as the preferred mode for solving all evaluation problems (see Baker, 1983 for a list).

Obviously, an analytical approach to evaluation design should be driven by what information is required by whom by when, by the credibility needed by the information analysts to do their job, and most importantly by the nature of the project or activity under review (Cronbach, 1980). Such precepts would suggest an eclectic approach, mixing journalistic, documentary, and effectiveness information as appropriate.

The information must provide policy feedback to the supporting agencies. This feature assumes that the funding source is either a contracting agency or an in-house manager. What kinds of policy feedback are appropriate? That depends in part on the nature of the formative evaluation team selected. Clearly, issues of project management might be a necessary concern. However, it is more likely that the substance to which the technology is directed, instruction, is a more useful area for feedback. At minimum, the formative evaluators should attend to the fidelity of the process by the project to the project's stated goals and procedures and to the kinds of contractual, monitoring, and other oversight arrangements that might be useful in the future. Furthermore, the evaluation report can consider specifically the features or tasks that might be included in the specification of future activities of the sort evaluated.

The tension of providing such information in a way that does not undo either the project activities under study or the receptivity of future projects to evaluation cannot be ignored. A fine line needs to be walked, keeping track of both the professional ethics applicable to contracting agency relationships i.e., (telling the truth) and to maintaining positive connections to the target R & D communities.

The information must provide timely and useful alternatives for the formative evaluation of the project(s) under study. This platitude takes serious effort to implement. It depends in no small measure in being informed accurately and intimately with the state of development of the project; and in the evaluation staff's sensitivity to the form as well as the substance of findings that might be useful to the project staff. This requirement also depends strongly on the level or stage of development of the technology activity. Early on, certain suggestions can be made and have potentially large effects. However, early on, the evidentiary base of such recommendations is likely to be weak. Later on, good evidence of project benefits and weaknesses can be more fully drawn; however, modification of the technology may be considerably less likely, and may cost more.

Thus, the model addresses macro or executive features of the development process rather than micro (or instructional) characteristics. Effectiveness data, based on careful assessment of an appropriate range of outcomes, constitute the critical feature of this model, for good management and good documentation have little importance when the question of "does it work?" is not well treated. We look toward a future in which such models will be routinely used and rational design and evaluation activities will actually drive instructional development, instead of seemingly evaporating following the approval of a new useful role, more realistic and practical than ever. It remains for the field to decide if it wishes to implement them, and how seriously.

Summary

We have tried to present in this chapter a discussion of outcome assessment that puts into context how measurement has evolved to its present state. We have attempted to detail the background of alternative viewpoints so that the reader can make informed professional decisions. We have also attempted to keep our eye on the ball of instruction, and urge those interested in outcome assessment not to get diverted by the intriguing, but occasionally irrelevant technical debates that suffuse the field of psychometrics. Good assessment depends more on hard thinking and good analysis than empirical solutions. It is for this reason, we advocate the use of criterion referenced measurement for the assessment of instructional technology outcomes, with the caveat that such measurement is difficult and must proceed beyond the often mindless way it is implemented at present.

Last, we believe that evaluation of technology outcomes is different from much of instructional assessment and that special attention to attributes of the assessment model are required.

REFERENCES

Aaron, H. (1977, October). Remarks before the Evaluation Research Society National Conference, Washington, D.C.

American Psychological Association. (1974). *Standards for educational and psychological tests*. Washington, D.C.: American Psychological Association.

Anderson, R.C. (1972). How to construct achievement tests to assess comprehension. *Review of Educational Research*, *42*, 140-170.

Baker, E.L. (1968). *Developing a researched based kindergarten reading program*. Inglewood, California: Southwest Regional Laboratory for Educational Research and Development (SWRL).

Baker, E.L. (1971). The effects of manipulated item writing constraints on the homogeneity of test items. *Journal of Educational Measurement*, 8(*4*), 305-309.

Baker, E.L. (1972). *Using measurement to improve instruction.* Paper presented at a Symposium of the Annual Meeting of the American Psychological Association, Honolulu.

Baker, E.L. (1982). The Specification of Writing Tasks. In A. Purves & S. Takala (Eds.), *Evaluation in Education: An Interaction Review Series*, *5*(3).

Baker, E.L., & Alkin, M.C. (1973). Formative evaluation in instructional development. *AV Communication Review*, *21*(4).

Baker, E.L. (1974). Formative evaluation of instruction. In J. Popham (Ed.), *Evaluation in education*, McCutchan.

Baker, E.L., & Saloutos, W.A. (1974). *Formative evaluation of instruction*. Los Angeles: UCLA Center for the Study of Evaluation.

Baker, E.L. (1978, January). *Is something better than nothing? Metaphysical test design*. Paper presented at the CSE Measurement and Methodology Conference, Los Angeles, CA.

Baker, E.L. (1980). Achievement tests in urban schools: New numbers. *CEMREL Monograph on Urban Education*, *4*.

Baker, E.L. (1983, October). *Evaluating educational quality: A rational design*. Invited paper, Educational Policy and Management, University of Oregon.

Baker, E.L. (1985, August). The impact of advance in artificial intelligence on test development. In the *Institutional Grant Proposal for NIE Center on Testing, Evaluation, and Standards: Assessing and Improving Educational Quality*. Los Angeles, CA: UCLA Center for the Study of Evaluation.

Baker, E.L., & Aschbacher, P. (1977). *Test design project*. Los Angeles, Ca.: Center for the Study of Evaluation.

Baker, E.L., Polin, L.G., Burry, J., & Walker, C. (1980, August). *Making, choosing and using tests: A practicum on domain-referenced testing*. Report to the National Institute of Education, Washington, D.C., (Grant No. OB-NIE-G-78-0213). Los Angeles: UCLA Center for the Study of Evaluation.

Baker, E.L., & Quellmalz, E.S. (1977). *Conceptual and design problems in competency based measurement. Long range plan, 1978-1982*. Los Angeles: UCLA Center for the Study of Evaluation.

Baker, E.L. Quellmalz, E., Enright, G. (1982). A Consideration of Topic Modality. Paper presented at a Symposium of the Annual Meeting of the American Educational Research Association, New York.

Baker, E.L., & Herman, J.L. (1983). Task Structure Design: Beyond Linkage, Journal of Educational Measurement, 20, 149-164.

Berk, R.A. (1980). A consumer's guide to criterion-referenced test "reliability". Paper presented at the annual meeting of the National Council on Measurement in Education, Boston.

Berk, R.A. (1980). Domain-referenced versus mastery conceptualization of criterion-referenced measurement: A clarification. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Block, J.H. (1971). Criterion-referenced measurements: Potential. School Review, 69, 289-298.

Bloom, B.S. (1968). Learning for mastery. Evaluation Comment, 1(2).

Bloom, B.S. (1969). Some theoretical issues relating to educational evaluation. In R. Tyler (Ed.), Educational evaluation: New roles, new means. The sixty-eighth yearbook for the National Society for the Study of Education, Part II, Chicago: National Society for the Study of Education, 26-50.

Bloom, B.S. (1980, November). Presentation made at the UCLA campus. University of California, Los Angeles.

Bloom, B.S., Englehart, M.D., Furst, E.J., Hill, W.H., & Krathwohl, D.R. (Eds.). (1956). Taxonomy of educational objectives: the classification of educational goals. Handbook I: Cognitive domain. New York: David McKay.

Bock, R.D., Mislevy, R.J., & Woodson, C. (1982). In Educational Researcher, 11, 4-11, 16.

Bock, R.D., Gibbons, R.D., & Muraki, E. (1985). Full-information item factor analysis. Chicago: NORC.

Bormuth, J.R. (1970). On a theory of achievement test items. Chicago, Il.: University of Chicago Press.

Brennan, R.L. (1974). Psychometric methods for criterion-referenced tests. University Awards Committee, State University of New York.

Brown, J.S., & Burton, R.R. (1984). Diagnostic models for procedural bugs in mathematics. Cognitive Science, 2, 155-192.

Bryk, A. (Ed.). (1983). Stakeholder-based evaluation. New Directions for Program Evaluation. Vol. 17. San Francisco: Jossey-Bass.

Buchanan, B.C. (1981). Research on Expert Systems. Report number CS-81-837, Computer Science Department, Stanford University.

Buros, O.K. (1977). Fifty years in testing: Some reminiscences, criticisms, and suggestions. Educational Researcher, 6, 9-15.

Carrol, J.B.A. (1963). A model school of learning. Teachers College Record, 64, 723-733.

Choppin, B.C. (1980, August). The IEA item banking project. Paper presented at the International Education Association Conference in Finland.

Clancy, W.J. (1982). Tutoring rules for guiding a case method dialogue. In D. Sleeman and J.S. Brown (Eds.), Intelligent tutoring systems. London: Academic Press.

Clark (1983, Winter). Reconsidering research on learning from media. Review of Educational Reserch, 53(4), 445-459.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfeld, F.D., & York, R.L. (1966). Equality of educational opportunity. Washington, D.C.: U.S. Government Printing Office.

Cronbach, L.J. (1971). Test validation. In R.L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D.C.: American Council on Education.

Cronbach, L.J., & Snow, R.F. (1977). Aptitudes and Instructional Methods. Irvington Publishing, Inc., New York.

Cronbach, L.J., et al. (1980). Toward reform of program evaluation. San Francisco: Jossey Bass.

Cronbach, L.J., Gleser, G.C., Nanda, R., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: John Wiley & Sons.

Cronbach, L.J., & Suppes, P. (Eds.). (1969). Research for tomorrow's schools -- disciplined inquiry for education. Report of the Committee on Educational Research of the National Academy of Education. London: Macmillan, Callien Macmillan, Ltd.

CSE Criterion-referenced Test Handbook. (1979). Los Angeles: UCLA Center for the Study of Evaluation.

Curtis, M.E., & Glaser, R. (1983). Reading theory and the assessment of reading achievement. Journal of Educational Measurement, 20, 133-147.

Danseresu, D.F., Rocklin, T.R., O'Donnell, A. M., Hythecker, Velma, I., Larson, C.O., Lambiotte, J.C., Young, M.D., & Flowers, L.F. Development and Evaluation of Computer-based Learning Strategy Training Modules. U.S. Army Research Institute for Behaviorial and Social Sciences, In Press.

Dehn, N. (1981). Story generation after tale-spin. Proceedings of the 7th International Joint Conference on Artificial Intelligence, Vancouver, British Columbia, Canada, 16-18.

Denham, C. (1975). Criterion-referenced, domain-referenced, and norm-referenced measurement: A parallax view. Educational Technology, 15(12), 9-13.

Doctorow, O. (1978). Some theoretical suggestions for a commutative test item operation. Unpublished manuscript.

Dunn, T.G., Lushene, R., & O'Neil, H.F., Jr. (1972). The complete automation of the Minnesota Multi-phasic Personality Inventory. Journal of Consulting and Clinical Psychology, 39, 381-387.

Dyer, M.G., & Lehner, W. (1982). Questioning answering for narrative memory. In J.F. Levy and W. Kingston (Eds.), Language and comprehension. New York: North Holland.

Dziuban, C.D., & Vickery, K.V. (1973). Criterion-referenced measurement: Some recent developments. Educational Leadership, 30(5), 483-486.

Ebel, R.L. (1971). Some limitations of criterion-reference measurement. Prepared for the annual meeting of the American Educational Research Association, Minneapolis, MN.

Ebel, R.L. (1972). Essentials of educational measurement. Englewood Cilffs, NJ: Prentice-Hall.

Ebel, R.L., & Anastasi, A. (1980). Abilities and the measurement of Achievement. In W. Schrader (Ed.), New directions for testing and measurement, measuring achievement: Progress over a decade. San Francisco, CA: Jossey Bass.

Emrick, J.A. (1971). An evaluation model for mastery testing. *Journal of Educational Measurement*, *8*, 321-326.

Engle, J.D., & Martuza, V.R. (1976, September). *A systematic approach to the construction of domain-referenced multiple-choice test items*. Paper presented at the annual meeting of the American Psychological Association, Washington, D.C.

Finn, P.J. (1978, March). *Generating domain-referenced, multiple choice test items from prose passages*. Paper presented at the annual meeting of the American Educational Research Association, Toronto.

Floden, R.E., Porter, A.C., Schmidt, W.H., & Freeman, D.J. (1980). Don't they all measure the same thing? Consequences of standardized test selection. In E. Baker and E. Quellmalz (Eds.) *Educational testing and evaluation, design, analysis, and policy*. Beverly Hills, CA: Sage Publications.

Frase, L.J. (1980). The demise of generality in measurement and research methodology. In E.L. Baker, & E.S. Quellmalz (Eds.), *Educational testing and evaluation: Design, analysis, and policy*. Beverly Hill, Ca.: Sage Publications.

Fredericksen, J.R., & Warren, B.M. (1985). *A cognitive framework for developing expertise in reading a research paper*. Cambridge, MA: Bolt, Berinek & Newman.

Freedle, R. (1985, June). *Implications of Language Programs in Artificial Intelligence for Testing Issues: Final Report Project 599-63*. Princeton,NJ: Educational Testing Services.

Fremer, J., & Anastasio, E.J. (1969). Computer-assisted item writing--I (Spelling items). Journal of Educational Measurement, 6(2), 69-74.

Gagne, R.M. (1965, 1977). The conditions of learning (1st & 3rd ed.). New York: Holt, Reinhart and Windston.

Gagne, R.M. (1977). Analyses of lectures. In L. Briggs (Ed.), Instructional design: Principles and applications. Englewood Cliffs, NJ: Educational Technology Publications.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 18, 519-21.

Glaser, R., & Nitko, A.J. (1971). Measurement in learning and instruction. In R.L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D.C.: American Council on Education.

Guttman, L. (1969). Integration of test design and analysis. In Proceedings of the 1969 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service.

Haladyna, T., & Roid, G. (1977). An empirical comparison of three approaches to achievement testing. Paper presented at the annual meeting of the American Psychological Association, San Francisco.

Haladyna, T., & Roid, G. (1978). The role of instructional sensitivity in the empirical review of criterion-referenced tests. Mcmouth, Or.: Teaching Research.

Hambleton, R.K. (1978). On the use of cut-off scores with criterion-referenced tests in instructional settings. Journal of Educational Measurement, 15(4), 277-290.

Hambleton, R.K. (1980). Test Score Validity and Cut-off Scores in R. Berk (Ed.), *Criterion-Referenced testing: State of the art*. Baltimore: The Johns Hopkins University Press.

Hambleton, R.K., & Simon, R. (1980). *Steps for constructing criterion-referenced tests*. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Hambleton, R.K., Swaminathan, H., Algina, J., & Coulson, D.B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research*, *48*, 1-47.

Haney, W. (1979). Trouble over testing. *Educational Leadership*, 37(8), 640-650.

Haney, W., & Madaus, G. (1978). Making sense of the competency testing movement. *Harvard Educational Review*, *48*(4), 462-484.

Harnischfeger, A., & Wiley, D. (1975). *Achievement Test Score Decline: Do We Need to Worry*? Chicago, IL: CEMREL, Inc.

Harris, C.W. (1962). *Measurement of change*. Milwaukee, WI: University of Wisconsin Press.

Harris, C.W. (1972). An interpretation of Livingston's reliability coefficient for criterion-referenced tests. *Journal of Educational Measurement*, *9*, 27-29.

Harris, C.W. (1973). Problems of objectives-based measurement. In C.W. Harris, M.C., Alkin, & W.J. Popham (Eds.), *Problems in criterion-referenced measurement*. Los Angeles: UCLA Center for the Study of Evaluation.

Harris, C.W. (1980, July). Final report to National Institute of Education (Grant No. NIE-G-78-0085, Project No. 8-0244). Los Angeles, CA: UCLA Center for the Study of Evaluation, 2 Vols.

Harris, M.L., & Stewart, D.M. (1971). Application of classical strategies to criterion-referenced test construction: An example. Paper presented at the annual meeting of the American Educational Research Association, New York.

Harris, N.D.C. (1976). A course mapping technique. Instructional Science, 5, 153-180.

Hedl, J.J., Jr., O'Neil, H.F., Jr., & Hanson, D.N. (1973). The affective Reactions towards computer-based intelligence testing. Journal of Consulting and Clinical Psychology, 40, 217-222.

Hively, W. (1973). Introduction to domain-referenced testing. Educational Technology, 14, 5-10.

Hively, W. (1974). Domain referenced testing. Englewood Cliffs, NJ: Educational Testing Publications.

Hively, W., Maxwell, G., Rabehl, G., Sension, D., & Lundin, S. (1973). Domain-referenced curriculum evaluation: Technical handbook and a case study from the MINNEMAST Project. CSE Monograph Series in Evaluation, No. 1. Los Angeles: UCLA Center for the Study of Evaluation.

Hively, W., Patterson, J., & Page, S. (1968). A "universe defined" system of arithmetic achievement test. Journal of Educational Measurement, 5(4), 275-290.

Hoepfner, R., Stern, C., Nummedal, S.G., et al. (1971). *CSE-ECRC preschool/kindergarten test evaluations*. Los Angeles: UCLA Center for the Study of Evaluation.

Holland, J.G. & Skinner, B.F. (1961). *The analysis of behavior: A program for self-instruction*. New York: McGraw-Hill.

Hovland, C.I., Lumsdaine, A.A., & Sheffield, F.D. (1949). *Experiments on mass communication*. Princeton, New Jersey: Princeton University Press.

Hsu, T., & Carlson, M. (1973). Test construction aspects of the computer assisted testing model. *Educational Technology*, *13*(3), 26-27.

Ivens, S.H. (1970). *An investigation of item analysis, reliability and validity in relation to criterion-referenced tests*. Unpublished doctoral dissertation, Florida State University.

Johnson, W.L., & Soloway, E. (1983). *Proust: Knowledge-based program understanding* (Technical Report YaleU/CSD/RR#285). New Haven, CT: Yale University, Computer Science Department.

Keller, F.S. (1968). Goodbye, teacher... *Journal of Applied Behavior Analysis*, *1*, 78-89.

Kriewall, T. (1972). Aspects and applications of criterion-referenced tests. *Illinois School Research*, *9*(2), 5-21.

Landa, L.N., Kopstein, F.F., & Bennet, V. (1974). *Algorithmization in learning and instruction*. New Jersey: Educational Technology Publications.

Levine, M. (1976). The academic achievement test - its historical context and social functions. *American Psychologist*, *31*(3), 228-238.

Lindquist, E.F. (1970). The Iowa testing program: A retrospective view. Education, 81, 7-23.

Lindvall, C.M., & Cox, R.C. (1969). The role of evaluation in programs individualized instruction. In R.W. Tyler (Ed.), Educational evaluation: New roles, new means. The 68th yearbook of the National Society for the Study of Education, Part II. Chicago: National Society for the Study of Education.

Livingston, S.A. (1972). Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 9(1), 13-26.

Lumsdaine, A.A., & May, M.A. (1965). Mass communication and educational media. In P.R. Fransworth, O. McNemar, & Q. NcNemar (Eds.), Annual Review of Psychology. Palo Alto, CA: Annual Reviews, Inc., 16, 475-534.

Macready, G.B., & Merwin, J. (1973). Homogeneity within item forms in domain-referenced testing. Educational and Psychological Measurement, 33(2), 352-360.

Mager, R.F. (1962). Preparing instructional objectives. Palo Alto, CA: Feardon.

Markle, D.G. (1967). An exercise in the application of empirical methods to instructional systems design. Final report: The development of the Bell system first aid and personal safety course, American Institutes for Research, Palo Alto, CA. New York: American Telephone and Telegraph Co.

Markle, S.M. (1967). Empirical testing of programs. In P.C. Lange (Ed.), Programmed instruction. The Sixty-sixth Yearbook of the National Society for the Study of Education, Part II. Chicago: NSSE.

McClung, M.S. (1978). Are competency testing programs fair? Legal? Phi Delta Kappan, 59(6), 397-400.

Merril, D.M., & Tennyson, R.D. (1977). Teaching concepts: An instructional design guide. Englewood Cliffs, NJ: Educational Technology Publications.

Michigan State University. (1968). B-Step: A teacher education curriculum. Michigan State University (Xerox).

Millman, J. (1974). Criterion-referenced measurement. In W.J. Popham (Ed.), Evaluation in Education. Berkeley, CA: McCutchan Publishing Corp.

Millman, J. (1974). Sampling plan for domain-referenced tests. EducationalTechnology, 14(6), 17-21.

Millman, J., & Outlaw, W.S. (1977). Testing by computer. Ithaca, NY: Cornell University Extension Publications.

Montague, W.E., Ellis, J.A., & Wulfeck, W.H., II. (1983). Instructional quality inventory: A formative test for instructional development. Performance and Instruction Journal, Vol. 22(5).

Moore, N.K., Shoffer, M.T., & Seifert, R.F. (1985, January). Basic Skills Requirements for Selected Army Occupational Training Courses. Contemporary Educational Psychology, Vol 10(1).

National Education Association. (1977). Guidelines and cautions for considering criterion-referenced tests. Washington, DC: National Education Association.

Nifenecker, E.A. (1918). Bureaus of research in city school systems. In G. Whipple (Ed.), The measurement of educational products. The 17th yearbook of the National Society for the Study of Education, Part II. Bloomington, IL: Public School.

Nitko, A. (1973). Problems in the development of criterion-referenced tests: The IPI Pittsburgh experience. In C.W. Harris, M.C. Alkin, & W.J. Popham (Eds.), Problems in criterion-referenced measurement. Los Angeles, CA: UCLA Center for the Study of Evaluation.

O'Neil, H.F., Jr. (1979) (Ed.) Learning Strategies. New York: Academic Press.

O'Neil, H.F., Jr., & Richardson, F.C. (1980). Test anxiety and computer-based learning environments. In I. Sarason, (Ed.) Test anxiety, research and applications. Hillsdale, NJ. Lea/Wiley.

O'Neil, H.F., Jr., & Richardson, F.C. In Sieber, J.E., O'Neil, H.F., Jr., & Tobias, S. (Eds.), Anxiety and learning in computer-based learning environments: An overview. Anxiety, Learning, and Instruction, Lawrence Erlbaum Associates, Inc., Hillsdale, NJ.

Olympia, P.L., Jr. (1975). Computer generation of truly repeatable Examinations. Educational Technology, 15(6), 53-55.

Osburn, H.G. (1968). Item sampling for achievement testing. Education and Psychological Measurement, 28, 95-104.

Perrone, V. (1975). Alternatives to standardized testing. National Elementary Principal, 54(6), 96-101.

Pipho, C. (1978). Minimum competency testing in 1978: A look at state standards. Phi Delta Kappan, 59(9), 585-588.

Polin, L.G., & Baker, E.L. (1979). Qualitative analysis of test item attributes for domain-referenced content validity judgments. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Popham, W.J. (1975). Educational Evaluation. Englewood Cliffs, NJ: Prentice Hall.

Popham, W.J. (1978). Practical criterion-referenced measures for intrastate evaluation. Educational Technology, 18(5), 19-23.

Popham, W.J., & Baker, E.L. (1978). Rules for the development of instructional products. Inglewood, CA: Southwest Regional Laboratory for Educational Research and Development (SWRL).

Popham, W.J., & Baker, E.L. (1970). Systematic instruction. Englewood Cliffs, NJ: Prentice-Hall.

Popham, W.J., & Baker, E.L. (1973). Teacher competency development system. Englewood Cliffs, NJ: Prentice-Hall.

Popham, W.J., & Husek, T.R. (1969). Implications of criterion-referenced measurement. Journal of Educational Measurement, 6(1), 1-9.

Purves, A.C. et al. (1980, August). International study of achievement in written composition. Paper presented at the International Education Association Conference, Jyviskleya, Finland.

Quellmalz, E.S. (1980, June). Test design: Aligning specifications for assessment and instruction. Paper presented at the conference Evaluation in the 80's: Perspectives for the National Research Agenda. Los Angeles, CA: UCLA Center for the Study of evaluation.

Rankin, S. (1980). Detroit Public Schools' use of a test-triggered improvement stratregy. Presentation to the annual meeting of the American Educational Research Association, Boston.

Rice, J.M. (1897). The futility of the spelling grind I & II. Forum, 23, 163-172 & 409-419.

Roid, G.H., & Haladyna, T.M. (1978). A comparison of objective-based and modified Bormuth item writing techniques. Educational and Psychological Measurement, 38, 19-28.

Roid, G.H., Haladyna, T., & Shaughnessy, J. (1979). Item writing for domain-based test of prose learning. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Roid, G.H., Haladyna, T., & Shaughnessy, J. (1980). A comparison of item-writing methods for criterion-referenced tests. Paper presented at the National Council on Measurement in Education, Boston.

Rosen, M.J. (1968). An experimental design for comparing the effects of instructional media programming procedures: Subjective vs. objective revision procedures. Final report. Palo Alto, CA: American Institutes for Behavioral Sciences.

Rudner, L.M. (1978). A short and simple introduction to tailored testing. Paper presented at the annual meeting of the Eastern Educational Research Association, Williamsburg.

Sanders, J.R., & Murray, S. (1976). Alternatives for achievement testing. Educational Technology, 16(6), 17-23.

Schwartz, J.L., & Garet, M.S. (Eds.). (1982). Assessment in the Service of Instruction. Report to the Ford Foundation and the National Institute of Education. Cambridge Massachusetts Institute of Technology.

Scriven, M. (1967). Aspects of curriculum development. In R. Tyler (Ed.), Perspectives of curriculum evaluation. Chicago: Rand McNally.

Scriven, M. (1967). The methodology of evaluation. In R.W. Tyler, R.M. Gagne, & M. Scriven (Eds.), Perspectives of curriculum evaluation. AERA Monograph Series on Curriculum Evaluation, No. 1. Chicago: Rand McNally.

Simon, G.B. (1969). Comments on "Implications of criterion-referenced tests." Journal of Educational Measurement, 6, 259-260.

Skager, R. (1975). EPT material. Abstract of: Critical characteristics for differentiating among tests of educational achievement. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Skinner, B.F. (1958). Teaching machines. Science, 128, 969-977.

Smith, E.R., Tyler, R.W., & et al. (1942). Appraising and the recording student progress. New York: Harper & Brothers, The Progressive Education Association Publications.

Spearman, C. (1937). Psychology down the ages (Vol. 1). London: MacMillan.

Stenner, A.J., & Webster, W.J. (1971). Educational program audit handbook. Arlington, Virginia: The Institute for the Development of Educational Auditing.

Tienmann, P., Kroeker, L.P., & Markle, S.M. (1977). Teaching verbally-mediated coordinate concepts in an on-going college course. Paper presented at the annual meeting of the American Educational Research Association, New York.

Tienmann, P., & Markle, S.M. (1978). Analyzing instructional content: A guide to instruction and evaluation. Champaign, IL: Stipes Publishing.

Title IV, Elementary and Secondary Education Act (ESEA). (1965).

Tyler, R.W. (1943). *Constructing achievement tests*. Columbus, Ohio: Ohio State University.

Tyler, R.W. (1950). *Basic principles of curriculum and instruction*. Chicago: University of Chicago Press.

Tyler, R.W. (1951). The functions of measurement in improving instruction. In E.T. Linquist (Ed.), *Educational measurement*, Washington, D.C.: American Council on Education.

Tyler, R.W., & Sheldon, H.W. (1979, October). *Testing, Teaching and Learning: Report of a Conference on Research on Testing August 17-26, 1979*. Washington, D.C.: National Institute of Education.

Ward, J.G. (1980). Issues in testing: The perspective of organized teachers and professors. In R. Bossone (Ed.), *Proceedings: The Third National Conference of Testing: Uniting testing and teaching*. New York: Center for Advanced Study in Education.

Washburne, C. Winnetka. (1922). *School and Society*, 29, 37-50.

Weiner, B. (1979). A theory of motivation for some classroom experiences. *Journal of Educational Psychology*, *71*, 3-25.

Wirtz, W. (1978). *Report of the advisory panel on the SAT score decline*. New York: Office of Public Information, College Entrance Exam Board.

Wright, B.D. (1967). *Sample free test calibration and person measurement*. Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service.